

INTERNATIONAL SERIES IN PHYSICS

F. K. RICHTMYER, CONSULTING EDITOR



INTRODUCTION TO THEORETICAL PHYSICS

INTERNATIONAL SERIES IN PHYSICS

F. K. RICHTMYER, CONSULTING EDITOR



- Bacher and Goudsmit*—
ATOMIC ENERGY STATES
- Clark*—
APPLIED X-RAYS
- Condon and Morse*—
QUANTUM MECHANICS
- Davey*—
A STUDY OF CRYSTAL STRUCTURE
AND ITS APPLICATIONS
- Edwards*—
ANALYTIC AND VECTOR MECHANICS
- Eldridge*—
THE PHYSICAL BASIS OF THINGS
- Hardy and Perrin*—
THE PRINCIPLES OF OPTICS
- Harnwell and Livingood*—
EXPERIMENTAL ATOMIC PHYSICS
- Houston*—
PRINCIPLES OF MATHEMATICAL PHYSICS
- Hughes and DuBridge*—
PHOTOELECTRIC PHENOMENA
- Hund*—
HIGH-FREQUENCY MEASUREMENTS
- Koller*—
THE PHYSICS OF ELECTRON TUBES
- Pauling and Goudsmit*—
THE STRUCTURE OF LINE SPECTRA
- Richtmyer*—
INTRODUCTION TO MODERN PHYSICS
- Ruark and Urey*—
ATOMS, MOLECULES AND QUANTA
- Slater and Frank*—
INTRODUCTION TO THEORETICAL PHYSICS
- White*—
INTRODUCTION TO ATOMIC SPECTRA
- Williams*—
MAGNETIC PHENOMENA

INTRODUCTION TO THEORETICAL PHYSICS

BY

JOHN C. SLATER, Ph.D.

Professor of Physics, Massachusetts Institute of Technology

AND

NATHANIEL H. FRANK, Sc.D.

*Assistant Professor of Physics, Massachusetts
Institute of Technology*

FIRST EDITION
THIRD IMPRESSION

McGRAW-HILL BOOK COMPANY, Inc.

NEW YORK AND LONDON

1933

IIA Lib.



01178

COPYRIGHT, 1933, BY THE
MCGRAW-HILL BOOK COMPANY, INC.

PRINTED IN THE UNITED STATES OF AMERICA

*All rights reserved. This book, or
parts thereof, may not be reproduced
in any form without permission of
the publishers.*

PREFACE

The general plan of a book is often clearer if one knows how it came to be written. This book started from two separate sources. First, it originated in a year's lecture course of the same title, covering about the first two-thirds of the ground presented here, the part on classical physics. This course grew out of the conviction that the teaching of theoretical physics in a number of separate courses, as in mechanics, electromagnetic theory, potential theory, thermodynamics, tends to keep a student from seeing the unity of physics, and from appreciating the importance of applying principles developed for one branch of science to the problems of another. The second source of this book was a projected volume on the structure of matter, dealing principally with applications of modern atomic theory to the structure of atoms, molecules, and solids, and to chemical problems. As work progressed on this, it became evident that the structure of matter could not be treated without a thorough understanding of the principles of wave mechanics, and that such an understanding demanded a careful grounding in classical physics, in mechanics, wave motion, the theory of vibrating systems, potential theory, statistical mechanics, where many principles needed in the quantum theory are best introduced. The ideal solution seemed to be to combine the two projects, including the classical and the more modern parts of theoretical physics in a coherent whole, thus further increasing the unity of treatment of which we have spoken.

Two general principles have determined the order of presenting the material: mathematical difficulty, and order of historical development. Mechanics and problems of oscillations, involving ordinary differential equations and simple vector analysis, come first. Then follow vibrations and wave motion, introducing partial differential equations which can be solved by separation of variables, and Fourier series. Hydrodynamics, electromagnetic theory, and optics bring in more general partial differential equations, potential theory, and differential vector operations. Wave mechanics uses almost all the mathematical machinery which has been developed in the earlier part of the book. It is natural that the historical order is in general the

same as the order of increasing mathematical difficulty, for each branch of physics as it develops builds on the foundation of everything that has gone before. In cases where the two arrangements do not coincide, we have grouped together subjects of mathematical similarity, thus emphasizing the unity of which we have spoken.

In a book of such wide scope, it is inevitable that many important subjects are treated in a cursory manner. An effort has been made to present enough of the groundwork of each subject so that not only is further work facilitated, but also the position of these subjects in a more general scheme of physical thought is clearly shown. In spite of this, however, the student will of course make much use of other references, and we give a list of references, by no means exhaustive, but suggesting a few titles in each field which a student who has mastered the material of this book should be able to appreciate.

At the end of each chapter is a set of problems. The ability to work problems, in our opinion, is essential to a proper understanding of physics, and it is hoped that these problems will provide useful practice. At the same time, in many cases, the problems have been used to extend and amplify the discussion of the subject matter, where limitations of space made such discussion impossible in the text. The attempt has been made, though we are conscious of having fallen far short of succeeding in it, to carry each branch of the subject far enough so that definite calculations can be made with it. Thus a far surer mastery is attained than in a merely descriptive discussion.

Finally, we wish to remind the reader that the book is very definitely one on theoretical physics. Though at times descriptive material, and descriptions of experimental results, are included, it is in general assumed that the reader has a fair knowledge of experimental physics, of the grade generally covered in intermediate college courses. No doubt it is unfortunate, in view of the unity which we have stressed, to separate the theoretical side of the subject from the experimental in this way. This is particularly true when one remembers that the greatest difficulty which the student has in mastering theoretical physics comes in learning how to apply mathematics to a physical situation, how to formulate a problem mathematically, rather than in solving the problem when it is once formulated. We have tried wherever possible, in problems and text, to bridge the gap

between pure mathematics and experimental physics. But the only satisfactory answer to this difficulty is a broad training in which theoretical physics goes side by side with experimental physics and practical laboratory work. The same ability to overcome obstacles, the same ingenuity in devising one method of procedure when another fails, the same physical intuition leading one to perceive the answer to a problem through a mass of intervening detail, the same critical judgment leading one to distinguish right from wrong procedures, and to appraise results carefully on the ground of physical plausibility, are required in theoretical and in experimental physics. Leaks in vacuum systems or in electric circuits have their counterparts in the many disastrous things that can happen to equations. And it is often as hard to devise a mathematical system to deal with a difficult problem, without unjustifiable approximations and impossible complications, as it is to design apparatus for measuring a difficult quantity or detecting a new effect. These things cannot be taught. They come only from that combination of inherent insight and faithful practice which is necessary to the successful physicist. But half the battle is over if the student approaches theoretical physics, not as a set of mysterious formulas, or as a dull routine to be learned, but as a collection of methods, of tools, of apparatus, subject to the same sort of rules as other physical apparatus, and yielding physical results of great importance. The title of this book might have been aptly extended to "Introduction to the Methods of Theoretical Physics," for the aim has constantly been, not to teach a great collection of facts, but to teach mastery of the tools by which the facts have been discovered and by which future discoveries will be made.

In a subject about which so much has been written, it seems hardly practicable to acknowledge our indebtedness to any specific books. From many of those mentioned in the section on suggested references, and from many others, we have received ideas, though the material in general has been written without conscious following of earlier models. We wish to express thanks to several of our colleagues for suggestions, and particularly to Professors P. M. Morse and J. A. Stratton, who have read the manuscript with much care and have contributed greatly by their discussions.

CONTENTS

	PAGE
PREFACE.	v

CHAPTER I

POWER SERIES

INTRODUCTION.	1
1. POWER SERIES.	2
2. SMALL QUANTITIES OF VARIOUS ORDERS.	3
3. TAYLOR'S EXPANSION.	4
4. THE BINOMIAL THEOREM	4
5. EXPANSION ABOUT AN ARBITRARY POINT.	4
6. EXPANSION ABOUT A POLE.	5
7. CONVERGENCE.	5
PROBLEMS.	8

CHAPTER II

POWER SERIES METHOD FOR DIFFERENTIAL EQUATIONS

INTRODUCTION.	10
8. THE FALLING BODY.	11
9. FALLING BODY WITH VISCOSITY	11
10. PARTICULAR AND GENERAL SOLUTIONS FOR FALLING BODY WITH VISCOSITY.	14
11. ELECTRIC CIRCUIT CONTAINING RESISTANCE AND INDUCTANCE	16
PROBLEMS.	17

CHAPTER III

POWER SERIES AND EXPONENTIAL METHODS FOR SIMPLE HARMONIC VIBRATIONS

INTRODUCTION.	19
12. PARTICLE WITH LINEAR RESTORING FORCE.	19
13. OSCILLATING ELECTRIC CIRCUIT	20
14. THE EXPONENTIAL METHOD OF SOLUTION	21
15. COMPLEX EXPONENTIALS	22
16. COMPLEX NUMBERS.	23
17. APPLICATION OF COMPLEX NUMBERS TO VIBRATION PROBLEMS	25
PROBLEMS.	26

CHAPTER IV

DAMPED VIBRATIONS, FORCED VIBRATIONS, AND RESONANCE

INTRODUCTION.	27
18. DAMPED VIBRATIONAL MOTION.	27

	PAGE
19. DAMPED ELECTRICAL OSCILLATIONS.	28
20. INITIAL CONDITIONS FOR TRANSIENTS.	29
21. FORCED VIBRATIONS AND RESONANCE.	29
22. MECHANICAL RESONANCE.	30
23. ELECTRICAL RESONANCE.	31
24. SUPERPOSITION OF TRANSIENT AND FORCED MOTION.	33
25. MOTION UNDER GENERAL EXTERNAL FORCES.	35
26. GENERALIZATIONS REGARDING LINEAR DIFFERENTIAL EQUATIONS PROBLEMS.	36 37

CHAPTER V

ENERGY

INTRODUCTION.	39
27. MECHANICAL ENERGY.	40
28. USE OF THE POTENTIAL FOR DISCUSSING THE MOTION OF A SYSTEM.	42
29. THE ROLLING-BALL ANALOGY.	45
30. MOTION IN SEVERAL DIMENSIONS.	46
PROBLEMS.	46

CHAPTER VI

VECTOR FORCES AND POTENTIALS

INTRODUCTION.	48
31. VECTORS AND THEIR COMPONENTS.	48
32. SCALAR PRODUCT OF TWO VECTORS.	49
33. VECTOR PRODUCT OF TWO VECTORS.	50
34. VECTOR FIELDS.	51
35. THE ENERGY THEOREM IN THREE DIMENSIONS.	52
36. LINE INTEGRALS AND POTENTIAL ENERGY.	52
37. FORCE AS GRADIENT OF POTENTIAL.	53
38. EQUIPOTENTIAL SURFACES.	54
39. THE CURL AND THE CONDITION FOR A CONSERVATIVE SYSTEM.	55
40. THE SYMBOLIC VECTOR ∇	55
PROBLEMS.	56

CHAPTER VII

LAGRANGE'S EQUATIONS AND PLANETARY MOTION

INTRODUCTION.	58
41. LAGRANGE'S EQUATIONS.	58
42. PLANETARY MOTION.	60
43. ENERGY METHOD FOR RADIAL MOTION IN CENTRAL FIELD.	61
44. ORBITS IN CENTRAL MOTION.	62
45. JUSTIFICATION OF LAGRANGE'S METHOD.	64
PROBLEMS.	67

CHAPTER VIII

GENERALIZED MOMENTA AND HAMILTON'S EQUATIONS

INTRODUCTION.	69
46. GENERALIZED FORCES.	69

CONTENTS

xi

	PAGE
47. GENERALIZED MOMENTA	70
48. HAMILTON'S EQUATIONS OF MOTION	71
49. GENERAL PROOF OF HAMILTON'S EQUATIONS.	72
50. EXAMPLE OF HAMILTON'S EQUATIONS.	74
51. APPLICATIONS OF LAGRANGE'S AND HAMILTON'S EQUATIONS	75
PROBLEMS.	76

CHAPTER IX

PHASE SPACE AND THE GENERAL MOTION OF PARTICLES

INTRODUCTION.	79
52. THE PHASE SPACE	80
53. PHASE SPACE FOR THE LINEAR OSCILLATOR	81
54. PHASE SPACE FOR CENTRAL MOTION	82
55. NONCENTRAL TWO-DIMENSIONAL MOTION	83
56. CONFIGURATION SPACE AND MOMENTUM SPACE.	83
57. THE TWO-DIMENSIONAL OSCILLATOR	84
58. METHODS OF SOLUTION	86
59. CONTACT TRANSFORMATIONS AND ANGLE VARIABLES	87
60. METHODS OF SOLUTION FOR NONPERIODIC MOTIONS.	90
PROBLEMS.	90

CHAPTER X

THE MOTION OF RIGID BODIES

INTRODUCTION.	92
61. ELEMENTARY THEORY OF PRECESSING TOP	92
62. ANGULAR MOMENTUM, MOMENT OF INERTIA, AND KINETIC ENERGY	94
63. THE ELLIPSOID OF INERTIA; PRINCIPAL AXES OF INERTIA	95
64. THE EQUATIONS OF MOTION.	96
65. EULER'S EQUATIONS	98
66. TORQUE-FREE MOTION OF A SYMMETRIC RIGID BODY	98
67. EULER'S ANGLES.	100
68. GENERAL MOTION OF A SYMMETRICAL TOP UNDER GRAVITY	102
69. PRECESSION AND NUTATION	104
PROBLEMS.	105

CHAPTER XI

COUPLED SYSTEMS AND NORMAL COORDINATES

INTRODUCTION.	107
70. COUPLED OSCILLATORS	107
71. NORMAL COORDINATES	111
72. RELATION OF PROBLEM OF COUPLED SYSTEMS TO TWO-DIMENSIONAL OSCILLATOR.	114
73. THE GENERAL PROBLEM OF THE MOTION OF SEVERAL PARTICLES	117
PROBLEMS.	118

CHAPTER XII

THE VIBRATING STRING, AND FOURIER SERIES

INTRODUCTION.	120
74. DIFFERENTIAL EQUATION OF THE VIBRATING STRING	120

	PAGE
75. THE INITIAL CONDITIONS FOR THE STRING.	122
76. FOURIER SERIES	123
77. COEFFICIENTS OF FOURIER SERIES	124
78. CONVERGENCE OF FOURIER SERIES	125
79. SINE AND COSINE SERIES, WITH APPLICATION TO THE STRING	126
80. THE STRING AS A LIMITING PROBLEM OF VIBRATION OF PARTICLES	128
81. LAGRANGE'S EQUATIONS FOR THE WEIGHTED STRING	131
82. CONTINUOUS STRING AS LIMITING CASE.	131
PROBLEMS.	132

CHAPTER XIII

NORMAL COORDINATES AND THE VIBRATING STRING

INTRODUCTION.	134
83. NORMAL COORDINATES	134
84. NORMAL COORDINATES AND FUNCTION SPACE	137
85. FOURIER ANALYSIS IN FUNCTION SPACE.	139
86. EQUATIONS OF MOTION IN NORMAL COORDINATES.	140
87. THE VIBRATING STRING WITH FRICTION.	142
PROBLEMS.	144

CHAPTER XIV

THE STRING WITH VARIABLE TENSION AND DENSITY

INTRODUCTION.	146
88. DIFFERENTIAL EQUATION FOR THE VARIABLE STRING	146
89. APPROXIMATE SOLUTION FOR SLOWLY CHANGING DENSITY AND TENSION	147
90. PROGRESSIVE WAVES AND STANDING WAVES	149
91. ORTHOGONALITY OF NORMAL FUNCTIONS.	151
92. EXPANSION OF AN ARBITRARY FUNCTION USING NORMAL FUNC- TIONS.	152
93. PERTURBATION THEORY.	154
94. REFLECTION OF WAVES FROM A DISCONTINUITY	156
PROBLEMS.	158

CHAPTER XV

THE VIBRATING MEMBRANE

INTRODUCTION.	160
95. BOUNDARY CONDITIONS ON THE RECTANGULAR MEMBRANE	160
96. THE NODES IN A VIBRATING MEMBRANE.	162
97. INITIAL CONDITIONS	162
98. THE METHOD OF SEPARATION OF VARIABLES.	163
99. THE CIRCULAR MEMBRANE	164
100. THE LAPLACIAN IN POLAR COORDINATES.	164
101. SOLUTION OF THE DIFFERENTIAL EQUATION BY SEPARATION.	165
102. BOUNDARY CONDITIONS.	166
103. PHYSICAL NATURE OF THE SOLUTION	167
104. INITIAL CONDITION AT $t = 0$	168

CONTENTS

xiii

PAGE

105. PROOF OF ORTHOGONALITY OF THE J 's	169
PROBLEMS.	170

CHAPTER XVI

STRESSES, STRAINS, AND VIBRATIONS OF AN ELASTIC SOLID

INTRODUCTION.	172
106. STRESSES, BODY AND SURFACE FORCES	172
107. EXAMPLES OF STRESSES.	174
108. THE EQUATION OF MOTION	175
109. TRANSVERSE WAVES	176
110. LONGITUDINAL WAVES	178
111. GENERAL WAVE PROPAGATION.	179
112. STRAINS AND HOOKE'S LAW	180
113. YOUNG'S MODULUS.	182
PROBLEMS.	183

CHAPTER XVII

FLOW OF FLUIDS

INTRODUCTION.	185
114. VELOCITY, FLUX DENSITY, AND LINES OF FLOW	185
115. THE EQUATION OF CONTINUITY.	186
116. GAUSS'S THEOREM	187
117. LINES OF FLOW TO MEASURE RATE OF FLOW	188
118. IRROTATIONAL FLOW AND THE VELOCITY POTENTIAL	188
119. EULER'S EQUATIONS OF MOTION FOR IDEAL FLUIDS.	190
120. IRROTATIONAL FLOW AND BERNOULLI'S EQUATION	191
121. VISCOUS FLUIDS	192
122. POISEUILLE'S LAW	194
PROBLEMS.	195

CHAPTER XVIII

HEAT FLOW

INTRODUCTION.	197
123. DIFFERENTIAL EQUATION OF HEAT FLOW	197
124. THE STEADY FLOW OF HEAT.	198
125. FLOW VECTORS IN GENERALIZED COORDINATES.	199
126. GRADIENT IN GENERALIZED COORDINATES.	200
127. DIVERGENCE IN GENERALIZED COORDINATES.	200
128. LAPLACIAN	201
129. STEADY FLOW OF HEAT IN A SPHERE	201
130. SPHERICAL HARMONICS	202
131. FOURIER'S METHOD FOR THE TRANSIENT FLOW OF HEAT	203
132. INTEGRAL METHOD FOR HEAT FLOW	205
PROBLEMS.	209

CHAPTER XIX

ELECTROSTATICS, GREEN'S THEOREM, AND POTENTIAL THEORY

INTRODUCTION.	210
133. THE DIVERGENCE OF THE FIELD	210

	PAGE
134. THE POTENTIAL	211
135. ELECTROSTATIC PROBLEMS WITHOUT CONDUCTORS.	212
136. ELECTROSTATIC PROBLEMS WITH CONDUCTORS	215
137. GREEN'S THEOREM.	217
138. PROOF OF SOLUTION OF POISSON'S EQUATION.	217
139. SOLUTION OF POISSON'S EQUATION IN A FINITE REGION.	220
140. GREEN'S DISTRIBUTION.	221
141. GREEN'S METHOD OF SOLVING DIFFERENTIAL EQUATIONS.	222
PROBLEMS.	223

CHAPTER XX

MAGNETIC FIELDS, STOKES'S THEOREM, AND VECTOR
POTENTIAL

INTRODUCTION.	225
142. THE MAGNETIC FIELD OF CURRENTS	226
143. FIELD OF A STRAIGHT WIRE	228
144. STOKES'S THEOREM.	229
145. THE CURL IN CURVILINEAR COORDINATES.	229
146. APPLICATIONS OF STOKES'S THEOREM.	230
147. EXAMPLE: MAGNETIC FIELD IN A SOLENOID	231
148. THE VECTOR POTENTIAL	231
149. THE BIOT-SAVART LAW	233
PROBLEMS.	234

CHAPTER XXI

ELECTROMAGNETIC INDUCTION AND MAXWELL'S
EQUATIONS

INTRODUCTION.	235
150. THE DIFFERENTIAL EQUATION FOR ELECTROMAGNETIC INDUCTION	235
151. THE DISPLACEMENT CURRENT.	236
152. MAXWELL'S EQUATIONS.	239
153. THE VECTOR AND SCALAR POTENTIALS	241
PROBLEMS.	244

CHAPTER XXII

ENERGY IN THE ELECTROMAGNETIC FIELD

INTRODUCTION.	246
154. ENERGY IN A CONDENSER.	246
155. ENERGY IN THE ELECTRIC FIELD.	247
156. ENERGY IN A SOLENOID.	248
157. ENERGY DENSITY AND ENERGY FLOW.	249
158. POYNTING'S THEOREM.	250
159. THE NATURE OF AN E.M.F.	250
160. EXAMPLES OF POYNTING'S VECTOR	251
161. ENERGY IN A PLANE WAVE	253
162. PLANE WAVES IN METALS.	255
PROBLEMS.	256

CHAPTER XXIII

REFLECTION AND REFRACTION OF ELECTROMAGNETIC WAVES

	PAGE
INTRODUCTION.	258
163. BOUNDARY CONDITIONS AT A SURFACE OF DISCONTINUITY.	258
164. THE LAWS OF REFLECTION AND REFRACTION.	259
165. REFLECTION COEFFICIENT AT NORMAL INCIDENCE	260
166. FRESNEL'S EQUATIONS	262
167. THE POLARIZING ANGLE.	264
168. TOTAL REFLECTION.	265
169. THE OPTICAL BEHAVIOR OF METALS	267
PROBLEMS.	268

CHAPTER XXIV

ELECTRON THEORY AND DISPERSION

INTRODUCTION.	270
170. POLARIZATION AND DIELECTRIC CONSTANT.	271
171. THE RELATIONS OF P , E , AND D	273
172. POLARIZABILITY AND DIELECTRIC CONSTANT OF GASES	275
173. DISPERSION IN GASES.	275
174. DISPERSION OF SOLIDS AND LIQUIDS.	278
175. DISPERSION OF METALS.	280
PROBLEMS.	283

CHAPTER XXV

SPHERICAL ELECTROMAGNETIC WAVES

INTRODUCTION.	286
176. SPHERICAL SOLUTIONS OF THE WAVE EQUATION	286
177. SCALAR POTENTIAL FOR OSCILLATING DIPOLE	288
178. VECTOR POTENTIAL.	289
179. THE FIELDS.	290
180. THE HERTZ VECTOR	291
181. INTENSITY OF RADIATION FROM A DIPOLE	293
182. SCATTERING OF LIGHT.	293
183. POLARIZATION OF SCATTERED LIGHT.	295
184. COHERENCE AND INCOHERENCE OF LIGHT	295
185. COHERENCE AND THE SPECTRUM	298
186. COHERENCE OF DIFFERENT SOURCES	299
PROBLEMS.	299

CHAPTER XXVI

HUYGENS' PRINCIPLE AND GREEN'S THEOREM

INTRODUCTION.	302
187. THE RETARDED POTENTIALS.	303
188. MATHEMATICAL FORMULATION OF HUYGENS' PRINCIPLE.	305
189. APPLICATION TO OPTICS.	307
190. INTEGRATION FOR A SPHERICAL SURFACE BY FRESNEL'S ZONES	308

	PAGE
191. THE USE OF HUYGENS' PRINCIPLE	310
192. HUYGENS' PRINCIPLE FOR DIFFRACTION PROBLEMS.	310
193. QUALITATIVE DISCUSSION OF DIFFRACTION, USING FRESNEL'S ZONES	311
PROBLEMS.	314

CHAPTER XXVII

FRESNEL AND FRAUNHOFER DIFFRACTION

INTRODUCTION.	315
194. COMPARISON OF FRESNEL AND FRAUNHOFER DIFFRACTION.	315
195. FRESNEL DIFFRACTION FROM A SLIT	319
196. CORNU'S SPIRAL	320
197. FRAUNHOFER DIFFRACTION FROM RECTANGULAR SLIT.	323
198. THE CIRCULAR APERTURE.	324
199. RESOLVING POWER OF A LENS	325
200. DIFFRACTION FROM SEVERAL SLITS; THE DIFFRACTION GRATING PROBLEMS.	326 328

CHAPTER XXVIII

WAVES, RAYS, AND WAVE MECHANICS

INTRODUCTION.	329
201. THE QUANTUM HYPOTHESIS	330
202. THE STATISTICAL INTERPRETATION OF WAVE THEORY.	332
203. THE UNCERTAINTY PRINCIPLE FOR OPTICS.	333
204. WAVE MECHANICS	335
205. FREQUENCY AND WAVE LENGTH IN WAVE MECHANICS.	337
206. WAVE PACKETS AND THE UNCERTAINTY PRINCIPLE	337
207. FERMAT'S PRINCIPLE	339
208. THE MOTION OF PARTICLES AND THE PRINCIPLE OF LEAST ACTION PROBLEMS.	342 343

CHAPTER XXIX

SCHRÖDINGER'S EQUATION IN ONE DIMENSION

INTRODUCTION.	345
209. SCHRÖDINGER'S EQUATION.	345
210. ONE-DIMENSIONAL MOTION IN WAVE MECHANICS.	346
211. BOUNDARY CONDITIONS IN ONE-DIMENSIONAL MOTION	350
212. THE PENETRATION OF BARRIERS.	351
213. MOTION IN A FINITE REGION, AND THE QUANTUM CONDITION	353
214. MOTION IN TWO OR MORE FINITE REGIONS	355
PROBLEMS.	356

CHAPTER XXX

THE CORRESPONDENCE PRINCIPLE AND STATISTICAL
MECHANICS

INTRODUCTION.	358
215. THE QUANTUM CONDITION IN THE PHASE SPACE	358
216. ANGLE VARIABLES AND THE CORRESPONDENCE PRINCIPLE.	359

CONTENTS

xvii

PAGE

217. THE QUANTUM CONDITION FOR SEVERAL DEGREES OF FREEDOM	361
218. CLASSICAL STATISTICAL MECHANICS IN THE PHASE SPACE.	364
219. LIOUVILLE'S THEOREM	365
220. DISTRIBUTIONS INDEPENDENT OF TIME	366
221. THE MICROCANONICAL ENSEMBLE	367
222. THE CANONICAL ENSEMBLE	368
223. THE QUANTUM THEORY AND THE PHASE SPACE.	369
PROBLEMS.	371

CHAPTER XXXI

MATRICES

INTRODUCTION.	374
224. MEAN VALUE OF A FUNCTION OF COORDINATES.	374
225. PHYSICAL MEANING OF MATRIX COMPONENTS	375
226. INITIAL CONDITIONS, AND DETERMINATION OF c 'S.	377
227. MEAN VALUES OF FUNCTIONS OF MOMENTA	379
228. SCHRÖDINGER'S EQUATION INCLUDING THE TIME	381
229. SOME THEOREMS REGARDING MATRICES.	382
PROBLEMS.	384

CHAPTER XXXII

PERTURBATION THEORY

INTRODUCTION.	386
230. THE SECULAR EQUATION OF PERTURBATION THEORY	386
231. THE POWER SERIES SOLUTION	387
232. PERTURBATION THEORY FOR DEGENERATE SYSTEMS.	390
233. THE METHOD OF VARIATION OF CONSTANTS	391
234. EXTERNAL RADIATION FIELD.	392
235. EINSTEIN'S PROBABILITY COEFFICIENTS.	393
236. METHOD OF DERIVING THE PROBABILITY COEFFICIENTS.	395
237. APPLICATION OF PERTURBATION THEORY	396
238. SPONTANEOUS RADIATION AND COUPLED SYSTEMS.	399
239. APPLICATIONS OF COUPLED SYSTEMS TO RADIOACTIVITY AND ELECTRONIC COLLISIONS.	402
PROBLEMS.	404

CHAPTER XXXIII

THE HYDROGEN ATOM AND THE CENTRAL FIELD

INTRODUCTION.	406
240. THE ATOM AND ITS NUCLEUS	406
241. THE STRUCTURE OF HYDROGEN.	407
242. DISCUSSION OF THE FUNCTION OF r FOR HYDROGEN.	410
243. THE ANGULAR MOMENTUM	414
244. SERIES AND SELECTION PRINCIPLES.	416
245. THE GENERAL CENTRAL FIELD.	418
PROBLEMS.	423

CHAPTER XXXIV

ATOMIC STRUCTURE

	PAGE
INTRODUCTION.	421
246. THE PERIODIC TABLE.	421
247. THE METHOD OF SELF-CONSISTENT FIELDS.	430
248. EFFECTIVE NUCLEAR CHARGES.	431
249. THE MANY-BODY PROBLEM IN WAVE MECHANICS.	432
250. SCHRÖDINGER'S EQUATION AND EFFECTIVE NUCLEAR CHARGES.	432
251. IONIZATION POTENTIALS AND ONE-ELECTRON ENERGIES.	432
PROBLEMS.	437

CHAPTER XXXV

INTERATOMIC FORCES AND MOLECULAR STRUCTURE

INTRODUCTION.	439
252. IONIC FORCES	439
253. POLARIZATION FORCE.	439
254. VAN DER WAALS' FORCE	440
255. PENETRATION OR COULOMB FORCE	442
256. VALENCE ATTRACTION.	442
257. ATOMIC REPULSIONS	444
258. ANALYTICAL FORMULAS FOR VALENCE AND REPULSIVE FORCES.	444
259. TYPES OF SUBSTANCES: VALENCE COMPOUNDS	447
260. METALS.	449
261. IONIC COMPOUNDS	449
PROBLEMS.	451

CHAPTER XXXVI

EQUATION OF STATE OF GASES

INTRODUCTION.	454
262. GASES, LIQUIDS, AND SOLIDS.	454
263. THE CANONICAL ENSEMBLE	456
264. THE FREE ENERGY.	458
265. PROPERTIES OF PERFECT GASES ON CLASSICAL THEORY.	461
266. PROPERTIES OF IMPERFECT GASES ON CLASSICAL THEORY	462
267. VAN DER WAALS' EQUATION.	464
268. QUANTUM STATISTICS.	466
269. QUANTUM THEORY OF THE PERFECT GAS	468
PROBLEMS.	470

CHAPTER XXXVII

NUCLEAR VIBRATIONS IN MOLECULES AND SOLIDS

INTRODUCTION.	471
270. THE CRYSTAL AT ABSOLUTE ZERO	472
271. TEMPERATURE VIBRATIONS OF A CRYSTAL.	474
272. EQUATION OF STATE OF SOLIDS.	478
273. VIBRATIONS OF MOLECULES	480

CONTENTS

xix

	PAGE
274. DIATOMIC MOLECULES	481
275. SPECIFIC HEAT OF DIATOMIC MOLECULES	483
276. POLYATOMIC MOLECULES	485
PROBLEMS.	486

CHAPTER XXXVIII

COLLISIONS AND CHEMICAL REACTIONS

INTRODUCTION.	488
277. CHEMICAL REACTIONS.	488
278. COLLISIONS WITH ELECTRONIC EXCITATION	491
279. ELECTRONIC AND NUCLEAR ENERGY IN METALS	494
280. PERTURBATION METHOD FOR INTERACTION OF NUCLEI.	497
PROBLEMS.	499

CHAPTER XXXIX

ELECTRONIC INTERACTIONS

INTRODUCTION.	501
281. THE EXCLUSION PRINCIPLE	502
282. RESULTS OF ANTISYMMETRY OF WAVE FUNCTIONS	506
283. THE ELECTRON SPIN	507
284. ELECTRON SPINS AND MULTIPLICITY OF LEVELS	509
285. MULTIPLICITY AND THE EXCLUSION PRINCIPLE.	510
286. SPIN DEGENERACY FOR TWO ELECTRONS	512
287. EFFECT OF EXCLUSION PRINCIPLE AND SPIN	514
PROBLEMS.	516

CHAPTER XL

ELECTRONIC ENERGY OF ATOMS AND MOLECULES

INTRODUCTION.	518
288. ATOMIC ENERGY LEVELS	518
289. SPIN AND ORBITAL DEGENERACY IN ATOMIC MULTIPLETS	520
290. ENERGY LEVELS OF DIATOMIC MOLECULES.	522
291. HEITLER AND LONDON METHOD FOR H_2	523
292. THE METHOD OF MOLECULAR ORBITALS.	527
PROBLEMS.	530

CHAPTER XLI

FERMI STATISTICS AND METALLIC STRUCTURE

INTRODUCTION.	531
293. THE EXCLUSION PRINCIPLE FOR FREE ELECTRONS	531
294. MAXIMUM KINETIC ENERGY AND DENSITY OF ELECTRONS.	534
295. THE FERMI-THOMAS ATOMIC MODEL	535
296. ELECTRONS IN METALS	536
297. THE FERMI DISTRIBUTION.	540
PROBLEMS.	543

CHAPTER XLII

DISPERSION, DIELECTRICS, AND MAGNETISM

	PAGE
INTRODUCTION.	545
298. DISPERSION AND DISPERSION ELECTRONS	546
299. QUANTUM THEORY OF DISPERSION	548
300. POLARIZABILITY	549
301. VAN DER WAALS' FORCE	551
302. TYPES OF DIELECTRICS	553
303. THEORY OF DIPOLE ORIENTATION	554
304. MAGNETIC SUBSTANCES.	555
PROBLEMS.	558
SUGGESTED REFERENCES.	561
INDEX.	565

INTRODUCTION TO THEORETICAL PHYSICS

CHAPTER I POWER SERIES

The first result of a physical experiment is ordinarily a table of values, one column containing values of an independent variable, another of a dependent variable. In mechanics, the independent variable is ordinarily the time, the dependent variable the displacement. In thermodynamics, we may have two independent variables, as volume and temperature, and one dependent variable, the pressure. With electric currents, we may have the current flowing in some part of the circuit as dependent variable, the electromotive force applied as independent variable, as when in a vacuum tube we measure plate current as function of grid voltage. In electromagnetic theory, the electric or magnetic field strength, the dependent variable, is a function of four independent variables, the three coordinates of space, and time.

The relation between independent and dependent variable can be given by a table of values, by drawing a graph, or analytically by approximating the results by a mathematical formula. The last method is by far the most powerful, particularly if further calculations must be made using the experimental results, so that we are led to the study of mathematical functions. There are a good many well-known functions; for example, the algebraic functions, as $ax + bx^2$; the trigonometric functions, as $\sin(ax + b)$; exponential functions, as ae^{-bx} ; and rarer things like Bessel's functions, $J_n(x)$. It may be that, by inspection of the results, or for some theoretical reason, we may decide that some such well-known function can be used to describe our experimental data within the experimental error. But in actual physical

problems, we meet many functions which are not included among these well-known forms. The question presents itself, can we not get some general method of describing functions analytically, equally applicable to familiar and unfamiliar functions?

1. Power Series.—Power series present one such general method, on the whole the most useful one. The simplest form of power series is $A_0 + A_1x + A_2x^2 + \dots$, where the A 's are arbitrary coefficients. By giving these coefficients suitable values, we can make the series approach any desired function as closely as we please, with some exceptions as we shall note below. As examples of common series, we have first the polynomials (in which all A_n 's after a certain n are zero); and then many familiar infinite series, as

$$(1+x)^n = 1 + nx + \frac{n(n-1)}{2!}x^2 + \frac{n(n-1)(n-2)}{3!}x^3 + \dots, \quad (1)$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots, \quad (2)$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots, \quad (3)$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots \quad (4)$$

In fitting an experimental table of values, it is generally true that we cannot use one of these well-known series. We must determine coefficients to fit the data. A familiar process is that in which we know beforehand that the graph of the function should be a straight line. Then, either by actually plotting and estimating by means of a ruler, or by using least squares, we find the two constants of the linear relation $y = a + bx$. If the graph is slightly curved, we may be able to determine the constants of a parabola $y = a + bx + cx^2$ to fit it approximately. More complicated curves can be approximated by taking more terms. It is plain that, if there are n points determined experimentally, we can find a polynomial containing n coefficients which will just pass through them. But this is hardly a sportsmanlike thing to do, and generally we look for a function containing far fewer constants than the number of points we wish to fit. In other words, in practice, rather than using infinite series, we are accustomed to use only the first few terms of such a series.

2. Small Quantities of Various Orders.—The general justification of this method of using only a finite part of a series comes from considering small quantities of various orders, as they are called. A power series is practically useful only if it converges rather rapidly; that is, if each term is decidedly smaller than the one before it. If we imagine that a physical relation is really expressed by a rapidly converging infinite series, then the sum of all the terms after a certain one will be smaller than the inevitable errors of experiment, and may be neglected, leaving only a polynomial. Suppose, for instance, that the linear dimension d of a solid under pressure, expressed as a function of the pressure p , is given exactly by a series $d = d_0 - ap + bp^2 - \dots$. For small pressures, the change of length ap will be small compared with d_0 , and the second-order term bp^2 will be in turn small compared with ap (though of course this will not be true for much higher pressures, since ap will increase, and bp^2 will increase even more). We express this by saying that ap is small quantity of the first order, bp^2 a small quantity of the second order. It may well be that the second-order quantities are so small that we can neglect them, so that approximately $d = d_0 - ap$. Now if we are interested in finding the way in which the volume, proportional to d^3 , changes with pressure, we have accurately

$$d^3 = d_0^3 - 3d_0^2ap + (3d_0a^2 + 3d_0^2b)p^2 + \dots \quad (5)$$

But we are assuming that ap is small compared with d_0 , and bp^2 is small compared with ap , for all pressures used. Thus we readily see that the term in p^2 in this final expression (5) is small compared with the term in p , and can be neglected in comparison with the leading term d_0^3 , so that in d^3 , as in d , we can neglect the second order of small quantities. We could then have started with the abbreviated expression $d = d_0 - ap$, and have obtained the same result for d^3 , to the first order.

This method of cutting off infinite series at definite places, retaining only terms of a certain order, is very commonly used, and often is the only thing that simplifies computations with series enough to make them practically possible. But we must notice that the justification depends entirely on the physical situation, and can be different in different cases. Thus if we had to consider higher pressures in our problem above, we should have to retain the second-order terms, but perhaps could neglect

third-order ones. One must always use good physical judgment in neglecting small quantities. Now, of course, in many cases we do not need to neglect high powers at all. The problems which we meet will often have simple enough relations between the coefficients of the successive terms so that we can write down as many terms as we please, without trouble, as we can with the binomial or exponential series. But it always pays to inquire, if the high terms of the series get too complicated to work with successfully, if they cannot be neglected.

3. Taylor's Expansion.—We have been speaking of series representing functions obtained from experiment, or about which we do not have much information. But it may be that we have to work with a function whose analytical properties we know, and in that case there is a standard method of finding its series expansion, known as Taylor's theorem. This is as follows:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots, \quad (6)$$

where $f(x)$ is the function of x , $f(0)$ means the value of the function when $x = 0$, $f'(0)$ is the first derivative for $x = 0$, and so on, so that $f(x) = A_0 + A_1x + A_2x^2 + \cdots$, where $A_n = f^n(0)/n!$. To justify this, we need only differentiate n times, obtaining very easily

$$\begin{aligned} f^n(x) &= n(n-1) \cdots (2)(1)A_n + (n+1)(n) \cdots (2)A_{n+1}x \\ &\quad + (n+2)(n+1) \cdots (3)A_{n+2}x^2 + \cdots \\ &= n!A_n + \frac{(n+1)!}{1!}A_{n+1}x + \cdots \end{aligned}$$

If now we let $x = 0$, all terms but the first vanish, so that we have $f^n(0) = n!A_n$, or $A_n = f^n(0)/n!$.

4. The Binomial Theorem.—As an illustration of Taylor's expansion, we prove the binomial theorem, the expansion of $(1+x)^n$ given in Eq. (1). We have

$$\begin{aligned} f(x) &= (1+x)^n, \\ f'(x) &= n(1+x)^{n-1}, \\ f''(x) &= n(n-1)(1+x)^{n-2}, \end{aligned}$$

etc., by differentiation. Thus, setting $x = 0$, $(1+x)$ goes into 1, so that we have $f(0) = 1$, $f'(0) = n$, $f''(0) = n(n-1)$, etc., and $A_0 = 1$, $A_1 = n/1!$, $A_2 = n(n-1)/2!$, etc.

5. Expansion about an Arbitrary Point.—A slightly more general expansion is obtained by shifting the origin along the x axis to a point a . The expansion is

$$f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \dots \quad (7)$$

From Taylor's theorem, we can see immediately a general condition which a function must satisfy if it can be expanded in power series about a given point (by expanding about a point we mean setting up an expansion in powers of $x - a$, if a is the given point). The function and all its derivatives must be finite at the point in question, since otherwise some coefficients of the expansion will be infinite. Thus for example we cannot expand $1/x$ in power series in x : we have $f(0) = 1/0 = \text{infinite}$, and all the derivatives are also infinite. Such a point is called a singular point of the function. But by expanding about another point we can avoid this difficulty. Thus we can expand $1/x$ about a , if $a \neq 0$;

$f(a) = 1/a$, $f'(a) = -1/a^2$, $f''(a) = 1.2/a^3$, $f'''(a) = -1.2.3/a^4$, etc., so that

$$\frac{1}{x} = \frac{1}{a} - \frac{(x - a)}{a^2} + \frac{(x - a)^2}{a^3} - \frac{(x - a)^3}{a^4} + \dots \quad (8)$$

From this we can understand that a function can be expanded in power series about a point which is not a singular point.

6. Expansion about a Pole.—At some singular points, the function behaves like $1/x^n$, an inverse power of x . Such a singularity is called a pole. If $f(x)$ has a pole of order n at the origin, then by definition $x^n f(x)$ has no singularity at the origin, and can be expanded in power series $A_0 + A_1 x + \dots$. Thus we have for $f(x)$ the expansion

$$f(x) = \frac{A_0}{x^n} + \frac{A_1}{x^{n-1}} + \dots, \quad (9)$$

an infinite series starting with inverse powers, but turning into an ordinary series of positive powers after its n th term. A similar theorem holds for expansion about a pole at $x = a$. A singularity which is not a pole is called an essential singularity. An example of an essential singularity is that possessed by the function $e^{-1/x}$ at $x = 0$. This function approaches 0 as x approaches 0 through positive values, but becomes infinite as x approaches 0 through negative values, and no inverse power $1/x^n$ has such a behavior.

7. Convergence.—A series is said to converge if the process of adding its terms is one that can be carried out and that leads to a

definite answer. Thus $(1 - x)^{-1}$, by the binomial theorem, is equal to $1 + x + x^2 + x^3 + \dots$. Now if x is less than unity, and we try to add these terms, we get an answer. For example, if $x = 0.1$, we have $1 + 0.1 + 0.01 + 0.001 + 0.0001 + \dots = 1.111\dots$, which equals $(1 - 0.1)^{-1} = 1\frac{1}{9}$, as it should. But if x is greater than unity, this no longer holds: if $x = 2$, we have $1 + 2 + 4 + 8 + \dots$, which certainly is infinitely great, and leads to no definite value. Another situation is obtained if we set $x = -1$ in the series, when we have $1 - 1 + 1 - 1 + 1 - 1 + \dots$, a series which is said to oscillate (successive terms have opposite signs). As a matter of fact, we find that the series $1 + x + x^2 + x^3 + \dots$, which is called the geometric series, converges if x is between -1 and $+1$, but does not converge if x is equal to or greater than 1 , or equal to or less than -1 . This series illustrates two of the simplest types of nonconvergence of series, the simple divergence, in which terms get greater and greater, and the oscillation, where the terms have the same order of magnitude but alternate sign. There is still another type of series which does not converge, sometimes called the semiconvergent or asymptotic series, whose terms begin to decrease regularly as we go out in the series, but after a certain point start in increasing, and eventually become infinite. These asymptotic series often can be used for computation, for it can be shown in many cases that, if we retain terms just up to the smallest one, the resulting sum is a good approximation to the function the series is supposed to represent.

Our definition of convergence in the last paragraph was very crude. More exactly, a series converges if the sum of the first n terms approaches a limit as n increases indefinitely. This definition agrees with the usual procedure of the physicist, for he often computes by series, and he does it by adding a finite number of terms. He carries this far enough so that adding more terms does not change the sum, to the order of accuracy to which he works, which essentially means that the sum is approaching a limit.

To tell whether a given series converges is not always easy. In the first place, we can be sure in some cases that given Taylor's expansions cannot converge if the argument (that is, the independent variable), has too large a value. Thus $1 + x + x^2 + \dots$ does not converge if x is equal to, or greater than, 1 , and we could have seen this from the fact that the series equals $1/(1 - x)$

which has a singularity for $x = 1$ (being equal to $\frac{1}{0}$). Thus the function is infinite for $x = 1$, and the series to represent it could not converge. And increasing x beyond 1 cannot make the series converge again. In fact, as soon as the variable in a series becomes greater than the value for which the function has a singularity, the series will diverge. But it is a little more complicated than this would seem, for $1 + x + x^2 + \dots$ diverges also for x less than -1 , and there is no singularity here. As a matter of fact, a power series converges in general so long as the argument is less in absolute value than the smallest value for which there is a singularity, but not beyond. But this singularity can come from imaginary or complex values of the argument, so that we might well miss it completely if we did not consider imaginary values. For this reason, this criterion for convergence is rather tricky.

When we actually examine a series, we can often tell whether it converges or not. Surely a series cannot converge unless its successive terms get smaller and smaller. We can investigate this by the ratio test, taking the ratio of the n th term to the one before, and seeing how this ratio changes as we go out in the series. If the limiting ratio is less than 1, the series converges; if it is greater than 1, it diverges. If the ratio is just 1, the test gives no information. Thus for example with the series $x + x^2/2 + x^3/3 + \dots$, the ratio of the term in x^n to that in x^{n-1} is $\frac{x^n}{n} \frac{n-1}{x^{n-1}} = \frac{(n-1)}{n}x$. As n approaches infinity, $n-1$ and n become approximately equal, so that the ratio approaches x . Thus we see that if x is less numerically than unity, this series converges; if x is greater than unity, it diverges; if $x = 1$, we cannot say. From other information, we know that the series when $x = 1$, which is $1 + 1/2 + 1/3 + 1/4 + \dots$, diverges. But with the similar series $x + x^2/2^2 + x^3/3^2 + \dots$, where the ratio of terms also approaches x as we go out in the series, and the series again diverges for x greater numerically than unity, converges for x less than unity, we have just the other situation at $x = 1$: the series $1 + 1/2^2 + 1/3^2 + \dots$ converges.

Often a series can be approximately summed by comparison with an integral. Thus

$$1 + \frac{1}{2^n} + \frac{1}{3^n} + \dots = \sum_{z=1}^{\infty} \frac{1}{z^n} = \int_1^{\infty} \frac{dz}{z^n} \text{ approximately.}$$

The approximation is rather poor for the small values of z , but becomes better for large z values, on which the convergence depends. It would be a better approximation, for instance, to write $\frac{1}{10^n} + \frac{1}{11^n} + \cdots = \int_{10}^{\infty} \frac{dz}{z^n}$. From this we see that the series converges when $n > 1$, the integral being $\frac{-1}{(n-1)z^{n-1}}$ which is zero at the upper limit, but diverges if $n \leq 1$. For $n = 1$, for instance, the integral becomes logarithmically infinite at $z = \infty$.

Problems

1. Plot $-\frac{1}{x} + \frac{1}{x^2}$ as a function of x , and show that it has a minimum at $x = 2$. Expand in Taylor's series about this point, obtaining an expansion $y = A_0 + A_2(x-2)^2 + A_3(x-2)^3 + \cdots$, where necessarily the coefficient A_1 is zero. Now plot on the graph the successive approximations $y = A_0$, $y = A_0 + A_2(x-2)^2$, $y = A_0 + A_2(x-2)^2 + A_3(x-2)^3$, $y = A_0 + A_2(x-2)^2 + A_3(x-2)^3 + A_4(x-2)^4$, observing how they approximate the real curve more and more accurately.

2. *a.* Derive the series for the exponential, cosine and sine series, directly from Taylor's theorem.

b. Differentiate the series for $\sin x$ term by term, and show that the result is the series for $\cos x$.

3. In the series for e^x , set $x = 1$, obtaining a series for e . Using this series, compute the value of e to four decimal places.

4. Why does one always have series for $\ln(1+x)$ in powers of x , rather than for $\ln x$? From the series for $\ln(1+x)$, compute logarithms to base e of 1.1, 1.2, 1.3, 1.4, 1.5.

5. The function $1/(x-i)$, where $i = \sqrt{-1}$, has a singularity for $x = i$, but not for any real value of x . Show that nevertheless the series expansion about $x = 0$ diverges for x greater than 1 or less than -1 , obtaining the power series by Taylor's theorem, and separating real and imaginary parts of the series. This is an example of a case where the series diverges on account of singularities for complex values of x .

6. As a result of an experiment, we are given the table of values following:

x	y
1	7.0
2	11.1
3	15.2
4	19.3
5	23.2
6	27.1
7	30.8
8	34.5
9	38.2
10	41.7

Try to devise some practicable scheme for telling whether this function (in which, being a result of experiment, the values are only approximations), can be represented within the error of experiment by a linear, quadratic, cubic, etc., polynomial. Get the coefficients of the resulting series, and use them to find the value of the function and its slope at $x = 0$. Plot the points, the curve which approximates them, and the straight-line tangent to the curve at $x = 0$. It is legitimate to use graphical methods if you wish.

7. Expand $\tan^{-1} x$ in a power series about $x = 0$. Hints:

$$(a) \frac{d}{dx} \tan^{-1} (x) = \frac{1}{1+x^2}$$

$$(b) \frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \dots$$

$$(c) \int \frac{d}{dx} (\tan^{-1} x) dx = \tan^{-1} x + c.$$

What is the range of convergence of the resulting series? Calculate from this series the value of $\pi/4 = \tan^{-1} 1$ correct to 5 per cent. How many terms of the series are necessary to obtain this accuracy?

8. By a procedure analogous to that used in Prob. 7 expand $\sin^{-1} x$ in a power series about $x = 0$. Find the range of convergence for this series.

9. From the known Taylor's series for e^x , write the corresponding series for e^{-x^2} . By integrating this series obtain to 1 per cent a value for

$$\int_0^1 e^{-x^2} dx,$$

whose correct value is 0.748. . . .

10. Make use of the binomial theorem to obtain an expansion of $\sqrt{1+x}$ in ascending powers of $x^{1/2}$. What is the range of convergence?

11. Discuss by the ratio test the convergence of the following series:

$$(a) x + x^2/2 + x^3/3 + x^4/4 + \dots$$

$$(b) x + x^2/2^2 + x^3/3^2 + x^4/4^2 + \dots$$

$$(c) \text{The binomial expansion of } (1+x)^k, \text{ for nonintegral } k.$$

$$(d) \text{The series for } e^x, \sin x, \cos x.$$

CHAPTER II

POWER SERIES METHOD FOR DIFFERENTIAL EQUATIONS

Most important physical laws involve statements giving the relation between the rate of change of some quantity and other quantities. Such a relation, stated in mathematical language, is a differential equation—an equation containing derivatives of functions, as well as the functions themselves. For example, the fundamental law of mechanics is Newton's second law of motion: the force equals the time rate of change of the momentum. Or in electricity, in a circuit containing an inductance, the back electromotive force of the inductance equals a constant times the time rate of change of the current. But these differential relations are not in the form which can be used in making direct connection with experiment. One cannot directly plot graphs, or give tables of values, from them. One must rather solve the differential equations, that is, find algebraic relations between the variables, containing no differentiations, but consistent with the differential equations. For most of our course we shall be interested in finding such solutions of differential equations.

Solving differential equations is rather like integrating functions: there are no general rules. Individual cases must be treated by appropriate special methods. We shall meet some such special rules, and shall make much use of some of them. Those who have studied differential equations have learned a variety of such rules. But rather more important on the whole is a method which is applicable, though not always most convenient, in a very large number of cases: the method of power series. In general, the solution of a differential equation consists of a certain functional relation between variables. If we assume that this function is expanded in power series, our only problem is to determine the coefficients. And by substituting the series back into the differential equation, we can very often get conditions for determining them. We shall illustrate the method by examples.

8. The Falling Body.—Imagine a body moving vertically under the action of gravity. To describe its motion, we have an independent variable, the time t , and a dependent variable, the height x . Let the mass of the body be m , and let its velocity, which is of course dx/dt , be also called v . The force acting on it is F . Then Newton's law states that $F = \frac{d(mv)}{dt}$, where mv is the momentum. If the mass is constant (which does not always have to be the case, as we shall see in Prob. 7), we can rewrite the equation as $F = m dv/dt$, or $= ma$, where a is the acceleration. Substituting $v = dx/dt$, this is also $F = m d^2x/dt^2$. These are all forms of Newton's second law, written as differential equations. We shall first take the case where the force, like that of gravity on the earth's surface, is constant: $F = \text{constant} = -mg$, where g is the acceleration of gravity, and where the negative sign means that the force is downward. Then we have

$$F = -mg = m \frac{dv}{dt} = m \frac{d^2x}{dt^2}, \quad (1)$$

or $d^2x/dt^2 = dv/dt = -g$. These can be solved at once, by direct integration: integrating once with respect to t , $dx/dt = v = \text{constant} - gt = v_0 - gt$, where v_0 , the constant of integration, obviously means the value of the velocity when $t = 0$. Integrating again, and calling the second constant of integration x_0 , we have $x = x_0 + v_0 t - \frac{1}{2}gt^2$, containing now two arbitrary constants, the initial position and initial velocity. The presence of such arbitrary constants is the most characteristic feature of the solutions of differential equations. And we note that the number of arbitrary constants equals the number of integrations we must perform to get rid of the differentiations. If the differential equation is one of the first order (with only first derivatives in it), there will be one arbitrary constant in the solution; if it is of the second order (second derivatives), there will be two, and so on. And always the arbitrary constants must be determined so as to satisfy certain "initial conditions," such as the values of the position and velocity at $t = 0$.

9. Falling Body with Viscosity.—With the problem of the falling body, the solution has automatically come out as a polynomial in t , which is simply a power series that breaks off, so that there is no need of more complicated methods. But now let us take a more difficult case: we assume the body to be falling

through a viscous medium under the action of gravity. Here the force is a sum of two parts: gravity, $-mg$, and a frictional force depending on velocity. It is found experimentally that for small velocities this frictional force, in a viscous medium, is proportional to the velocity, with, of course, a negative coefficient, since it opposes the motion, changing signs with the velocity. Let it be called $-kv$, k being the coefficient, which depends in a complicated way on the shape and size of the body, and is proportional to the coefficient of viscosity of the fluid. Then we have

$$m \frac{dv}{dt} = -mg - kv,$$

or

$$m \frac{dv}{dt} + kv = -mg. \quad (2)$$

This is a simple sort of differential equation, in a standard form. It is

1. A linear differential equation. That is, it contains v and its derivatives (as v , dv/dt , d^2v/dt^2 , etc.) only in the first power (in dv/dt , kv), or the zero power ($-mg$, independent of v), not as squares or cubes [as, for example, $(dv/dt)^2$], or products (as $v dv/dt$).

2. A differential equation of the first order (containing no derivative higher than the first).

3. An inhomogeneous equation (it contains terms of both the first power and the zero power in v and its derivatives, while a homogeneous equation contains only terms of the same power, as all of the first power. That is, if the term $-mg$ were absent, the equation would be homogeneous).

We cannot solve Eq. (2) by direct integration, for if we integrate with respect to t , one term would be $\int v dt$, which we cannot evaluate, since v is an unknown function of time. Thus we must proceed differently. Let us assume that v is given by a power series in the time, $v = A_0 + A_1t + \dots$, and try to determine the coefficients. We do this by substituting the series in the equation. We have by direct differentiation

$$\frac{dv}{dt} = A_1 + 2A_2t + 3A_3t^2 + \dots + (n+1)A_{n+1}t^n + \dots$$

Then, substituting, we have

$$m[A_1 + 2A_2t + 3A_3t^2 + \dots + (n+1)A_{n+1}t^n + \dots] + k(A_0 + A_1t + A_2t^2 + \dots + A_nt^n + \dots) = -mg.$$

which we meet in the solution. We could compute from our series the value of v at any time t , knowing the initial velocity.

It happens in this case that we can recognize the infinite series as representing a familiar function. For we have

$$e^{-\frac{k}{m}t} = 1 - \frac{k}{m}t + \frac{1}{2!} \frac{k^2}{m^2}t^2 - \frac{1}{3!} \frac{k^3}{m^3}t^3 + \dots,$$

which has close connection with our series, so that we can write at once

$$\begin{aligned} v &= A_0 + \left(\frac{k}{m}A_0 + g \right) \frac{\left(e^{-\frac{k}{m}t} - 1 \right)}{k/m} \\ &= \left(A_0 + \frac{mg}{k} \right) e^{-\frac{k}{m}t} - \frac{mg}{k}. \end{aligned} \quad (7)$$

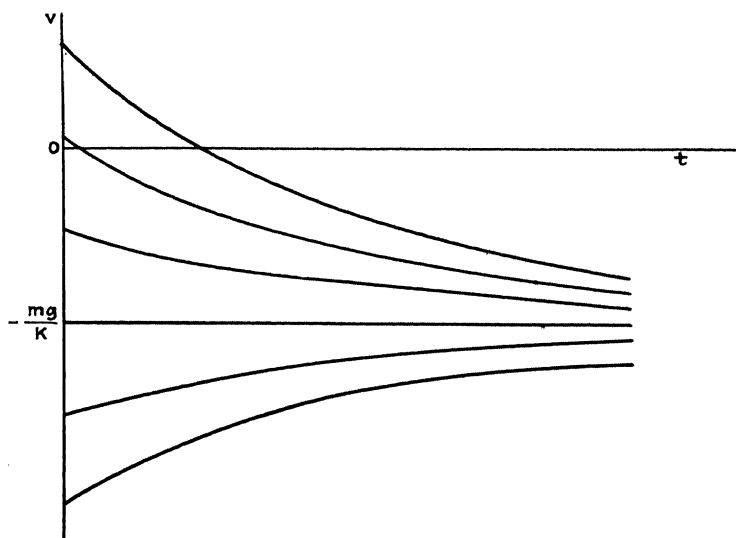


FIG. 1.—Velocity of damped falling body, with various initial conditions.

We can see the physical properties of the solution most clearly from the graph in Fig. 1. No matter what the initial velocity may have been, the particle finally settles down to motion with a constant speed, given by $-mg/k$. The initial velocity is A_0 , and if this is greater than the final velocity, the body slows down; if it is less, it speeds up, to attain this final speed.

10. Particular and General Solutions for Falling Body with Viscosity.—It is instructive to notice that we can solve our

problem in an elementary way. Our equation is $mdv/dt + kv = -mg$. Plainly a particular solution is given by assuming a constant velocity. Then dv/dt is zero, so that the equation is $, or $v = -mg/k$. But this is not the most general solution, for it does not have an arbitrary constant; it represents merely the particular case in which the initial velocity happened to be just the correct final value, and is unable to describe any other initial condition. To get a general solution, we proceed as follows: we take the homogeneous equation $mdv/dt + kv = 0$, which we obtain from our inhomogeneous equation by leaving out the term $-mg$. We can easily solve this: writing it $dv/v = -(k/m)dt$, and integrating, we have $\ln v = -(k/m)t + \text{constant}$, and taking the exponential, $v = \text{constant} \times e^{-(k/m)t}$, where the constant is arbitrary. Then the sum of this general solution of the homogeneous equation, and the particular solution $-(m/k)g$ of the inhomogeneous equation, is the solution we desire. We may prove this easily. For we have$

$$\begin{aligned} \left(m \frac{d}{dt} + k\right) \left(-\frac{m}{k}g\right) &= -mg \\ \left(m \frac{d}{dt} + k\right) \left(Ce^{-\frac{k}{m}t}\right) &= 0. \end{aligned}$$

Adding,

$$\left(m \frac{d}{dt} + k\right) \left(Ce^{-\frac{k}{m}t} - \frac{m}{k}g\right) = -mg,$$

showing that the function $Ce^{-(k/m)t} - (m/k)g$ satisfies the differential Eq. (2).

The procedure we have just used is an illustration of the general rule: *A general solution of an inhomogeneous equation is obtained by adding a particular solution of the inhomogeneous equation, and a general solution of the related homogeneous equation.* In this statement, the terms "particular solution" and "general solution" are used in a technical sense: a "particular solution" is one which satisfies the differential equation but has no arbitrary constants; a "general solution" is one which has its full complement of arbitrary constants. The proof of the rule in general is carried out just as in our case, adding the particular solution of the inhomogeneous equation and a general solution of the homogeneous equation, and showing that the sum satisfies the inhomogeneous equation. One thing should be noted: the properties we have been discussing depend entirely on the linear character

of the differential equation, for it is only with linear functions f that $f(x_1) + f(x_2) = f(x_1 + x_2)$.

11. Electric Circuit Containing Resistance and Inductance.—

The theory of the electrical circuit reminds one in many ways of mechanical principles: electric current is analogous to velocity, charge to displacement, electromotive force to mechanical force. Thus in a circuit containing a resistance, inductance, and condenser, all in series, the current can flow through the circuit, piling up in the condenser because it cannot flow through. Let q be the charge on one plate of the condenser ($-q$ being the charge on the other), and let i be the current flowing through the circuit toward the condenser plate in question, so that the current measures just the amount of charge per second flowing onto the condenser plate, or $i = dq/dt$ (as $v = dx/dt$). Now let the coefficient of self-induction of the circuit be L , the resistance R , the capacity of the condenser C . Then there are three e.m.fs. (electromotive forces) acting on the current, in addition to a possible external e.m.f. E from a battery: the back e.m.fs. of induction, resistance, and capacity. The first is $-L \frac{di}{dt}$, the electromotive force induced in a circuit when the current changes; the second is $-Ri$, the value familiar from Ohm's law; the third is $-q/C$, as given by the elementary law of the condenser. These are all negative, for they act to oppose the current. Now the law of the circuit is that the total e.m.f. acting on the circuit is zero:

$$-L \frac{di}{dt} - Ri - \frac{q}{C} + E = 0,$$

or

$$L \frac{di}{dt} + Ri + \frac{q}{C} = E. \quad (8)$$

This is a differential equation. Let us take the special case where there is no condenser, so that the equation is $L di/dt + Ri = E$. The equation is then exactly analogous to the equation $mdv/dt + kv = F$, which we had for a falling body with viscosity. And we see that self-induction is analogous to inertia, resistance to viscosity. The analogy is often valuable.

If now the applied e.m.f. E of the battery is constant, the problem can be solved mathematically just as before, and we find $i = \text{constant} \times e^{-(R/L)t} + E/R$. The first term is the transient

effect, of arbitrary size, as we see from the arbitrary constant, rapidly dying out as time goes on, while the second is the constant value given by Ohm's law, the value to which the current tends if we wait long enough.

Problems

1. Show that the solution $v = (A_0 + mg/k)e^{-(k/m)t} - mg/k$ reduces properly to uniformly accelerated motion in the limiting case where the viscous resistance vanishes. Illustrate this graphically, showing curves for several different k 's, and finally for $k = 0$, all with the same initial velocity.

2. A raindrop weighs 0.1 gm., and after falling from rest reaches a limiting speed of 1,000 cm. per second by the time it reaches the earth. How long did it take to reach half its final speed? Nine tenths of its final speed? How far did it travel before reaching half its final speed? For how long could its velocity be described by the simple law $v = -gt$ to an error of 1 per cent?

3. At high velocities, the viscous resistance is proportional to the third power of the velocity. Assuming this law, set up the differential equation for a particle falling under gravity and acted on by such a viscous drag. Solve by power series, obtaining at least four terms in the expansion for v as a function of t . Draw graphs of velocity as function of time, and discuss the solutions physically.

4. Using the same law of viscosity as in the preceding problem, but assuming no gravitational force, solve by direct integration of the differential equation for the case of a particle starting with given initial velocity and being damped down to rest. Show by Taylor's expansion of this function that it agrees with the special case of the power series of the preceding problem obtained by letting the gravitational force be zero.

5. A large coil has a resistance of 0.7 ohm, inductance of 5 henries. Until $t = 0$, no current is flowing in the coil. At that moment, a battery of 5 volts e.m.f. is connected to it. After 5 sec., the battery is short-circuited and the current in the coil allowed to die down. Compute the current as function of the time, drawing a curve to represent it.

6. A coil having $L = 10$ henries, $R = 1$ ohm, has no current flowing in it until $t = 0$. Then it has an applied voltage increasing linearly with the time, from zero at $t = 0$, to 1 volt at $t = 1$ sec. After $t = 1$, the e.m.f. remains equal to 1 volt. By series methods find the current at any time, and plot the curve.

7. Suppose we have a rocket, shot off with initial velocity v_0 , and thereafter losing mass according to the law $m = m_0(1 - ct)$, where m is the mass at any time, m_0 the initial mass at $t = 0$, c is a constant, and where the mass lost does not have appreciable velocity after it leaves the rocket. Show that on account of the loss of mass the rocket is accelerated, just as if a force were acting on a body of constant mass. The rocket is acted on by a viscous resisting force in addition. Taking account of these forces, find the differential equation for its velocity as a function of time, and integrate the equation directly. Now find also the solution for v as a power series in the time. Show that the resulting series agrees with that obtained by expanding the

exact solution. Calculate the limiting ratio of successive terms in the power series, as we go out in the series, and from this result obtain the region of convergence of the series. Is this result reasonable physically? What happens in the exact solution outside the range of convergence?

8. In a radioactive disintegration, the number of atoms disintegrating per second, and turning into atoms of another sort, is simply proportional to the total number of radioactive atoms present. Write down the differential equation for the number of atoms present at any time, and find its solution. Assuming that half the atoms of a sample of radium disintegrate in 1,300 years, how many would decay in the first year?

9. If at the same time radium were being produced at a constant rate by disintegration of uranium, how would this change the situation in the preceding problem? Set up the new differential equation. Assuming that we start without any radium, but with pure uranium, find the amount of radium as a function of the time. Show that the amount of radium approaches an equilibrium amount, which it reaches in time, whether the initial amount of radium is greater or less than the equilibrium amount.

10. Find a series solution for the differential equation $m dv/dt + kv = c/t$, where c is a constant, representing a damped motion under the action of an external force which decreases inversely proportionally to the time, the series having the form $v = a_1/t + a_2/t^2 + \dots$. Show that this series is divergent for all values of t . Show that the differential equation is formally satisfied by the expression $v = e^{-t} \int_{-\infty}^t \frac{e^t}{t} dt$. This solution is convergent for t

negative. The integral $\int_{-\infty}^t \frac{e^t}{t} dt$ is known as the exponential integral function, and is important in physics and mathematics. It is frequently calculated by using the above divergent series. Explain how this procedure might be valid.

11. Suppose a particle is acted on by a damping force proportional to the velocity, and to a force which varies sinusoidally with the time. Solve the resulting differential equation for velocity as function of time, by the series method, by expanding the force in power series in the time. Can you recognize the analytical form of the resulting power series?

12. Solve by power series Bessel's equation $\frac{d^2y}{dx^2} + \frac{1}{x} \frac{dy}{dx} + y = 0$. The result is Bessel's function of the zero order, $J_0(x)$. From the series, plot $J_0(x)$ for x between 0 and 5.

13. The equation for Bessel's function of the m th order, $J_m(x)$, is $\frac{d^2y}{dx^2} + \frac{1}{x} \frac{dy}{dx} + \left(1 - \frac{m^2}{x^2}\right)y = 0$. Solve by power series, showing that the first term in the expansion is that in x^m . Plot $J_1(x)$ for x between 0 and 5. Bessel's functions oscillate, like the sine and cosine, all the way to infinity. We shall use them in discussing standing waves in a circular membrane, and for many other problems. The second independent solution of the equation is infinite at the origin, and hence cannot be expanded in power series.

CHAPTER III

POWER SERIES AND EXPONENTIAL METHODS FOR SIMPLE HARMONIC VIBRATIONS

In the last chapter we have found a general method of power series for solving differential equations, and have applied it to the problem of motion under viscous forces. Next we consider the same method, applied to somewhat different problems: a particle acted on by restoring forces proportional to the distance, or an electric circuit containing inductance and capacity.

12. Particle with Linear Restoring Force.—Suppose that the force acting on a particle is proportional to the displacement from a fixed position, and opposite to the displacement, a so-called linear restoring force. This force is $-kx$, if x is the displacement, k a constant. For the moment we assume that there is no gravitational or other external force acting. Then the equation of motion is $m d^2x/dt^2 = -kx$, or

$$m \frac{d^2x}{dt^2} + kx = 0 \quad (1)$$

This is a homogeneous linear differential equation of the second order, with constant coefficients (that is, m , k are independent of time). We solve it in series as before. If $x = A_0 + A_1t + A_2t^2 + \dots$, we have immediately, by the method used before, $(2mA_2 + kA_0) + (3 \cdot 2mA_3 + kA_1)t + (4 \cdot 3mA_4 + kA_2)t^2 + \dots = 0$.

Thus, setting the separate coefficients equal to zero, and solving one equation after the other, we find

$$\begin{aligned} A_2 &= -\frac{1}{2} \frac{k}{m} A_0, & A_3 &= -\frac{1}{3!} \frac{k}{m} A_1, \\ A_4 &= \frac{1}{4!} \left(\frac{k}{m} \right)^2 A_0, & A_5 &= \frac{1}{5!} \left(\frac{k}{m} \right)^2 A_1, \\ &\dots\dots\dots \end{aligned} \quad (2)$$

These equations determine all the coefficients in terms of two arbitrary ones, A_0 and A_1 , which are the two arbitrary constants

to be expected in the solution of a second-order differential equation. The solution may be written

$$x = A_0 \left[1 - \frac{1}{2!} \frac{k}{m} t^2 + \frac{1}{4!} \left(\frac{k}{m} \right)^2 t^4 - \dots \right] + A_1 \left[t - \frac{1}{3!} \frac{k}{m} t^3 + \frac{1}{5!} \left(\frac{k}{m} \right)^2 t^5 - \dots \right]. \quad (3)$$

We now observe that these series represent well-known functions: the first is the cosine, the second the sine, except for a factor, so that we have

$$x = A_0 \cos \sqrt{k/m} t + A_1 \sqrt{m/k} \sin \sqrt{k/m} t. \quad (4)$$

Thus the motion is a periodic one, as shown by the sinusoidal functions. The period T is found from the fact that when t increases by T , the sine or cosine must come back to its initial value, which it does when its argument (that is, the thing whose cosine we are taking), increases by 2π . Thus $\sqrt{k/m} T = 2\pi$, $T = 2\pi\sqrt{m/k}$, the familiar formula for the period in simple harmonic motion. From this, the frequency ν is given by $\nu = 1/T = (1/2\pi)\sqrt{k/m}$, and the angular velocity ω by $\omega = 2\pi\nu = \sqrt{k/m}$. It is often convenient to use these relations in rewriting the equation of motion, writing it

$$d^2x/dt^2 + \omega^2 x = 0, \text{ or } d^2x/dt^2 + 4\pi^2\nu^2 x = 0. \quad (5)$$

13. Oscillating Electric Circuit.—In the last chapter, we have seen that the equation for an electric circuit containing resistance, inductance, and capacity, is $L di/dt + Ri + q/C = E$, where i is the current, q the charge on the condenser, and E the impressed electromotive force. We also saw that $i = dq/dt$. Substituting, we obtain

$$L \frac{d^2q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = E. \quad (6)$$

This is an inhomogeneous second-order linear differential equation for q , which becomes homogeneous if $E = 0$. We consider that case, and in particular let R be zero. Then the problem becomes mathematically equivalent to the preceding one, and has the differential equation $d^2q/dt^2 + q/LC = 0$. The solution is $q = A_0 \cos \sqrt{1/LC} t + A_1 \sqrt{LC} \sin \sqrt{1/LC} t$, so that the current oscillates in the circuit. By differentiating, we can find the current directly instead of the charge: $i = dq/dt =$

$-A_0\sqrt{1/LC} \sin \sqrt{1/LC} t + A_1 \cos \sqrt{1/LC} t$, so that the oscillations of charge and current are similar. The period of oscillation is given by $T = 2\pi\sqrt{LC}$, increasing as either the inductance or the capacity becomes large.

14. The Exponential Method of Solution.—We have found that the solutions of our vibration problems, as well as of several other differential equations, come out either as exponential functions, or as sines or cosines. As a matter of fact, any homogeneous linear differential equation, with constant coefficients, has such solutions. On account of the importance of this type of equation, we shall consider its solution specially. Let us take a second-order differential equation,

$$\frac{d^2y}{dx^2} + a\frac{dy}{dx} + by = 0, \quad (7)$$

a type which includes the mechanical and electrical problems we have worked with. We can show very easily that this has an exponential solution, $y = e^{kx}$. For let us substitute this function into the equation. We have $dy/dx = ky$, $d^2y/dx^2 = k^2y$, so that the equation becomes $(k^2 + ak + b)e^{kx} = 0$. This equation is factored, and since e^{kx} is not always zero, the other factor must be, and we have $k^2 + ak + b = 0$, or solving the quadratic by formula, $k = -a/2 \pm \sqrt{(a/2)^2 - b}$. Thus if k equals either $k_1 = -a/2 + \sqrt{(a/2)^2 - b}$, or $k_2 = -a/2 - \sqrt{(a/2)^2 - b}$, e^{kx} is a solution of the equation. We have, in fact, two independent solutions.

Now if we have two independent solutions of a second-order linear homogeneous differential equation, we can readily show that any linear combination of them is itself a solution. If such a solution has two arbitrary constants, it is a general solution. Thus we can write the general solution of Eq. (7)

$$y = Ae^{k_1x} + Be^{k_2x},$$

or

$$y = e^{-(a/2)x} [Ae^{\sqrt{(a/2)^2 - b}x} + Be^{-\sqrt{(a/2)^2 - b}x}]. \quad (8)$$

This is the solution, with its two arbitrary constants, and it might seem as if no further discussion were necessary. But there is an interesting feature still to consider: the quantity $(a/2)^2 - b$ under the radical may easily be negative, and the square root imaginary, so that we have to investigate the exponentials of imaginary quantities.

Suppose, for example, that the damping term is zero: $a = 0$, and the differential equation is $d^2y/dx^2 + by = 0$. This is the only case we have so far worked out in detail. Then the solution becomes $y = Ae^{i\sqrt{b}x} + Be^{-i\sqrt{b}x}$, where $i = \sqrt{-1}$. But we have already seen that the solution of this same equation is $C \cos \sqrt{b}x + D \sin \sqrt{b}x$. If both forms are right, there must be connections between exponential and sinusoidal functions, which we now proceed to investigate.

15. Complex Exponentials.—Let us investigate the function e^{ix} by series methods. We have at once

$$\begin{aligned} e^{ix} &= 1 + ix - \frac{x^2}{2!} - \frac{ix^3}{3!} + \frac{x^4}{4!} \cdots \\ &= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} \cdots\right) + i\left(x - \frac{x^3}{3!} + \cdots\right), \end{aligned}$$

or

$$e^{ix} = \cos x + i \sin x.$$

Similarly we have

$$e^{-ix} = \cos x - i \sin x. \quad (9)$$

We can solve for $\cos x$ by adding these equations and dividing by 2, or for $\sin x$ by subtracting and dividing by $2i$:

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}. \quad (10)$$

These theorems are fundamental in the study of exponential and sinusoidal functions.

In terms of the formulas of the last paragraph, we can readily see that our two formulations of simple harmonic motion are both correct. For we have

$$\begin{aligned} Ae^{i\sqrt{b}x} + Be^{-i\sqrt{b}x} \\ &= A (\cos \sqrt{b}x + i \sin \sqrt{b}x) + B (\cos \sqrt{b}x - i \sin \sqrt{b}x) \\ &= (A + B) \cos \sqrt{b}x + i(A - B) \sin \sqrt{b}x, \end{aligned}$$

or one constant times the cosine plus another times the sine, which is the more familiar solution. By giving A and B suitable complex values, we can have both coefficients real. But to know how to do this, and to understand the whole process, we should study complex numbers for themselves. Let us then make a little survey of the theory of complex numbers.

16. Complex Numbers.—A complex number is usually written $A + Bi$, where A and B are real, $i = \sqrt{-1}$. It is often plotted in a diagram: we let abscissas represent real parts of numbers, ordinates the imaginary parts, so that A measures the abscissa, B the ordinate, of the point representing $A + Bi$. Every point in the plane corresponds to a complex number, and *vice versa*. All real numbers lie along the axis of abscissas, all pure imaginaries along the axis of ordinates, and the other complex numbers between. But it is also often convenient to think of a complex

number as being represented, not merely by a point, but by the vector from the origin out to the point. The fundamental reason for this is that these vectors obey the parallelogram law of addition, just as force or velocity vectors do (see Fig. 2). The vector treatment is suggestive in many ways.

For example, we can consider the angle between two complex numbers. Thus, any real number, and any pure imaginary number, are at an angle of 90 deg. to each other. Or, the number $1 + i$ is at an angle of 45 deg. with either 1 or i . When a complex number is regarded as a vector, we can

describe it by two quantities: the absolute magnitude of the vector, or its length, $\sqrt{A^2 + B^2}$; and the angle which it makes with the real axis, or $\tan^{-1} B/A$.

The vector representation of complex numbers has very close connection with complex exponential functions. Let us consider the complex number $e^{i\theta}$, where θ is a real quantity. As we have seen, this equals $\cos \theta + i \sin \theta$, so that the real part is $\cos \theta$, the imaginary part $\sin \theta$. The vector representing this number is then a vector of unit magnitude, for $\sqrt{\cos^2 \theta + \sin^2 \theta} = 1$. Further, it makes just the angle θ with the real axis. We can see interesting special cases. The number $e^{\pi i/2} = i$, as we can see at once from the vector diagram, or from the fact that it equals $\cos \pi/2 + i \sin \pi/2 = i$. Similarly $e^{\pi i} = -1$, $e^{2\pi i} = e^{4\pi i} = \dots = 1$. This last result shows that the exponential

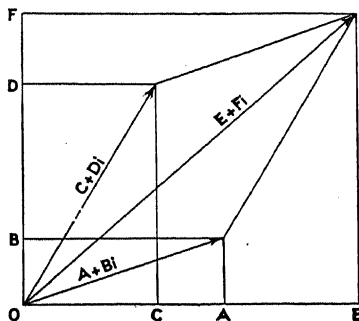


FIG. 2.—Law of addition of complex vectors. The vector $E + Fi$ represents the vector sum of $A + Bi$ and $C + Di$. Evidently $OE = OA + AE = OA + OC$, and $OF = OB + OD$. Hence $E + Fi = (A + C) + (B + D)i$.

function of an imaginary argument is periodic with period $2\pi i$, similarly to the sine and cosine of a real argument.

Next we look at the number $re^{i\theta}$, where r , θ are both real. It differs from $e^{i\theta}$ in that both real and imaginary parts are multiplied by the same real factor r , which simply increases the length of the vector to r , without changing the angle. Thus $re^{i\theta}$ is a vector of length r , angle θ . As a result, we can easily write any complex number in complex exponential form: $A + Bi = re^{i\theta}$,

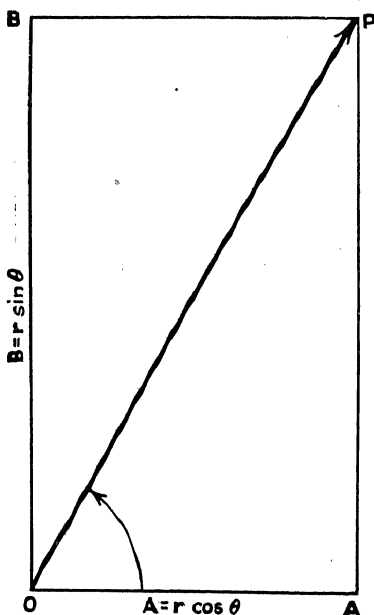


FIG. 3.—The complex number P equals either $A + Bi$, or $re^{i\theta}$.

where $r = \sqrt{A^2 + B^2}$, $\theta = \tan^{-1} B/A$, or $A = r \cos \theta$, $B = r \sin \theta$ (see Fig. 3). We may use these results in showing what happens when two complex numbers are multiplied together. Suppose we wish to form the product $(A + Bi)(C + Di)$. Of course, multiplying directly, this equals $(AC - BD) + (AD + BC)i$, so that we can easily find real and imaginary parts of the product, but this is not very informing. It is better to write $A + Bi = r_1 e^{i\theta_1}$, $C + Di = r_2 e^{i\theta_2}$. Then the product is $(r_1 e^{i\theta_1})(r_2 e^{i\theta_2}) = (r_1 r_2) e^{i(\theta_1 + \theta_2)}$. That is, the magnitude of the product of two complex numbers is the product of the magnitudes, and the angle is the sum of their angles.

Suppose we have a complex number $re^{i\theta}$, and consider the closely related number $re^{-i\theta}$. The second is called the conjugate of the first. If we have a complex number in the form $A + Bi$, its conjugate is $A - Bi$. Or in general, if we change the sign of i wherever it appears in a complex number, we obtain its conjugate. Graphically, the vector representing the conjugate of a number is the mirror image of the vector representing the number itself, in the axis of real numbers. Now conjugate numbers have two important properties: the sum of a number and its conjugate is real (for the imaginary parts just cancel in taking this sum), and the product is real (for this equals

$r^2 e^{i(\theta-\theta)} = r^2$). The second fact is useful in finding the absolute magnitude of a complex number: if z is complex, \bar{z} its conjugate (this is the usual notation), then $\sqrt{z\bar{z}}$ equals the absolute magnitude of z . From the other fact, we may find the real and imaginary parts of complex numbers: $\frac{z + \bar{z}}{2}$ equals the real part

of z , and as we can easily show, $\frac{z - \bar{z}}{2i}$ equals the imaginary part. We see examples in our relations between sinusoidal and exponential functions where e^{-ix} is the conjugate of e^{ix} , so that $\frac{e^{ix} + e^{-ix}}{2}$ should, and does, equal the real part of e^{ix} , or $\cos x$, and $\frac{e^{ix} - e^{-ix}}{2i}$ equals the imaginary part, or $\sin x$.

17. Application of Complex Numbers to Vibration Problems.—

There are two different, though related, ways of applying complex numbers to vibration problems. The first, and perhaps more logical, is directly suggested by what we have done. We found for undamped vibrations that $y = Ae^{i\sqrt{b}x} + Be^{-i\sqrt{b}x}$. Now naturally we wish y to be real, since it represents a real displacement. To do this, we make use of the proposition that we have just found, that the sum of a complex number and its conjugate is real. Since $e^{-i\sqrt{b}x}$ is the conjugate of $e^{i\sqrt{b}x}$, we achieve the desired result if we make $B = \bar{A}$, for then the whole second term is just the conjugate of the first. Incidentally, if we write $A = \frac{C}{2} e^{-i\alpha}$, we have

$$y = \frac{C}{2} e^{i(\sqrt{b}x - \alpha)} + \frac{C}{2} e^{-i(\sqrt{b}x - \alpha)} = C \cos(\sqrt{b}x - \alpha), \quad (11)$$

giving a form, in terms of amplitude C and phase α , which is often useful and important.

The second method of treatment is more common, particularly in electrical applications. Suppose we work directly with the complex solution $y = Ae^{i\sqrt{b}x}$, but consider that only the real part is of physical significance. This real part, as we have seen, is half the sum of this quantity and its conjugate, so that, except for a factor of 2, it comes to the same thing we have considered before. However, it is often easier to think of it in this way, and the process of using a complex solution, and finally taking the real part, is very common. Of course, if A is real,

the real part is simply $A \cos \sqrt{b}x$; if A is complex, we may write it $Ce^{-i\alpha}$, and the real part of the product is $C \cos (\sqrt{b}x - \alpha)$. This second method is particularly interesting in discussing simple harmonic motion, where x is replaced by t , and \sqrt{b} by ω , so that we are considering the real part of $Ae^{i\omega t}$. The complex number is given by a vector of length A , rotating in the complex plane with angular velocity ω . And its real part is simply the projection of the vector along the real axis. Thus it corresponds exactly to the most elementary formulation of simple harmonic motion, as the projection of a circular motion on a diameter.

Problems

1. Show directly that the solution $A \sin \omega t + B \cos \omega t$ for the particle moving with simple harmonic motion can also be written $C \cos (\omega t - \alpha)$. Find C and α as functions of A and B , and *vice versa*. The constant C is called the amplitude of the motion, and α is called the phase. Note that α can be regarded as an angle, measured in radians.

2. A pendulum 1 m. long is held at an angle of 1 deg. to the vertical, and released with an initial velocity of 5 cm. per second toward the position of equilibrium. Find amplitude and phase of the resulting motion.

3. A circuit contains resistance, inductance, and capacity, but there is no impressed e.m.f. Solve the differential equation in series, and show by comparison of the first few terms that the series represents the function $e^{-(R/2L)t}(A \sin \omega t + B \cos \omega t)$, where $\omega^2 = 1/LC - R^2/4L^2$.

4. In an oscillatory circuit, show that the phases of the charge and the current differ by 90 deg.

5. Given a complex number represented by a vector, what is the nature of the vector representing its square root; its cube root? Find the three cube roots of unity, the four fourth roots, the five fifth roots, plotting them in the complex plane, and giving real and imaginary components of each. With one of the cube roots, in terms of its real and imaginary parts, cube by direct multiplication and show that the result is unity.

6. Find real and imaginary parts of $\sqrt{A + Bi}$, $\frac{1}{A + Bi}$, $\frac{1}{\sqrt{A + Bi}}$, where A, B are real.

7. Show that $\ln(-a) = \pi i + \ln a$, or $3\pi i + \ln a$, or in general $n\pi i + \ln a$, where n is an odd integer.

8. Prove that if we have a complex solution of the problem of a vibrating particle, the real part of this complex function is itself a solution of the problem.

9. Show that in general a linear homogeneous differential equation of the n th order with constant coefficients has n independent exponential solutions of the sort we have considered.

10. Show that if we have n independent solutions of an n th order differential equation, then an arbitrary linear combination of these solutions, containing n coefficients, is a general solution of the equation.

CHAPTER IV

DAMPED VIBRATIONS, FORCED VIBRATIONS, AND RESONANCE

We have now reached the point where we can discuss a wide range of problems in oscillatory mechanical or electrical systems. The general question we shall take up is that of a system containing inertia, damping force proportional to the velocity, and restoring force proportional to the displacement, under the action of an impressed force. This leads to an inhomogeneous second-order linear differential equation, of the form

$$m \frac{d^2x}{dt^2} + 2mk \frac{dx}{dt} + m\omega^2 x = F(t), \quad (1)$$

where the coefficients $2mk$ and $m\omega^2$ of the damping and restoring force terms, respectively, are written in this way to obtain a simple result. The term $F(t)$, which makes the equation inhomogeneous, is the impressed force, a function of time. The solution of such an inhomogeneous equation, as we have seen, can be written as a sum of two parts. One is a particular solution of the problem, the so-called forced motion, a steady-state solution which persists as long as the force is applied. The other is the transient term, a general solution of the corresponding homogeneous equation obtained by setting $F = 0$. This transient proves to be a damped simple harmonic motion, an oscillation whose amplitude decreases exponentially with time, soon passing away, and leaving only the steady-state solution. The amplitude and phase of the transient are determined so that the whole motion will have the correct initial displacement and velocity, its two arbitrary constants being chosen to fit the initial conditions.

18. Damped Vibrational Motion.—We first consider the transient motion, whose equation is obtained from (1) above by setting $F = 0$. In the preceding chapter we have seen that the solution can be written

$$x = e^{-kt}(Ae^{\sqrt{k^2 - \omega^2}t} + Be^{-\sqrt{k^2 - \omega^2}t}). \quad (2)$$

There are three cases: (1) $k^2 - \omega^2 < 0$; (2) $k^2 - \omega^2 = 0$; (3) $k^2 -$

$\omega^2 > 0$. The first is the case where the damping is small. Here $\sqrt{k^2 - \omega^2} = i\sqrt{\omega^2 - k^2}$, and the radical is real. Then we have the same sort of expression we have considered before, and to get a real answer we must write $B = \bar{A}$, or else we can take the real part of a complex quantity. Let us do the latter: the solution is the real part of $Ae^{-kt}e^{i\sqrt{\omega^2 - k^2}t}$, or is $Ce^{-kt} \cos(\sqrt{\omega^2 - k^2}t - \alpha)$. This is like a simple harmonic motion, of angular velocity $\sqrt{\omega^2 - k^2}$, phase α , but with an amplitude Ce^{-kt} which continually decreases with time, and it is called damped simple harmonic motion. For small damping, the angular velocity can be expanded in power series, and is $\omega - \frac{k^2}{2\omega} \dots$, differing from ω

by a small quantity of the second order. Thus, for example, a pendulum which is slightly damped will have its period only very slightly altered by the damping. The amplitudes of successive swings go down in exponential fashion, on account of the factor e^{-kt} . Thus the logarithms of the amplitudes go down linearly with the time, and as a result this kind of damping is known as logarithmic damping. The decrease in the logarithm of the amplitude in a period is known as the logarithmic decrement.

The other extreme case is the third, where $k^2 - \omega^2 > 0$, and there is nothing complex about the solution at all. It simply consists of two exponential terms, with only real coefficients. The resulting motion is not oscillatory, but merely damps down gradually to zero. The limiting case, $k^2 - \omega^2 = 0$, is called the critical case, and is most easily discussed as the limit of either of the others. An interesting practical application of all the cases is found in the problem of the vibrations of galvanometers. A galvanometer without damping oscillates back and forth with simple harmonic motion. With slight damping, it has nearly the same frequency, but a logarithmic decrement. As the damping is made greater and greater, the period gets larger and larger, until finally at critical damping and beyond there are no oscillations at all. The galvanometer, if displaced, simply settles slowly back to its normal position.

19. Damped Electrical Oscillations.—The corresponding electrical problem is given by the circuit containing resistance, inductance, and capacity, and the equation is

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{q}{C} = 0. \quad (3)$$

The solution is

$$q = Ce^{-(R/2L)t} \cos(\omega t - \alpha), \quad (4)$$

where

$$\omega = \sqrt{1/LC - R^2/4L^2}.$$

This is the same solution which we found in Prob. 3 of the last chapter by the series method. It is an interesting illustration of the simplicity of the exponential method of solving the equation. As we see, the current oscillates with an angular velocity which, for small R , differs only slightly from the undamped angular velocity $\sqrt{1/LC}$, but it has a logarithmic damping, which is greater the greater R is.

20. Initial Conditions for Transients.—To fix the two arbitrary constants of the transient, we must fit the initial displacement and velocity. Thus, for instance, consider the solution in the form

$$x = Ce^{-kt} \cos(\sqrt{\omega^2 - k^2} t - \alpha).$$

Assume that at $t = 0$, $x = x_0$, and $dx/dt = v_0$. From the first,

$$x_0 = C \cos \alpha. \quad (5)$$

To apply the second, we have

$$\begin{aligned} \frac{dx}{dt} &= -Ce^{-kt} \sqrt{\omega^2 - k^2} \sin(\sqrt{\omega^2 - k^2} t - \alpha) \\ &\quad - kCe^{-kt} \cos(\sqrt{\omega^2 - k^2} t - \alpha). \end{aligned}$$

Thus

$$v_0 = C\sqrt{\omega^2 - k^2} \sin \alpha - kC \cos \alpha. \quad (6)$$

By simultaneous solution of Eqs. (5) and (6) we can find C and α in terms of x_0 and v_0 .

21. Forced Vibrations and Resonance.—Our next task is to find a particular solution of Eq. (1) containing the external applied force. To do this, we shall first solve the case where the force is a sinusoidal function of the time, a very important special case. This leads to a solution also sinusoidal with the same frequency, with an amplitude proportional to the amplitude of the force, but for which the constant of proportionality depends on the frequency, becoming large out of all proportion if the impressed frequency is nearly equal to the natural frequency. This phenomenon of enormously exaggerated response of the

oscillating system to a certain impressed frequency is called resonance; it is of great physical importance.

Familiar examples of resonance will occur to one. In mechanics, it is well known that a pendulum can be set swinging with large oscillations if it receives small periodic impulses, timed to synchronize with its own period, whereas any other impressed frequency would soon get out of step with the oscillations it sets up, and would force them to die down again. Acoustical resonance is illustrated by the way in which one vibrating tuning fork will set another into vibration if both have the same pitch, but not otherwise. Another acoustical example comes from Helmholtz's resonators: air chambers vibrating with a definite pitch, which are set into resonant vibration if sound of that particular pitch falls on them, but not appreciably by any other pitch, so that they can be used to pick out a particular note in a complicated sound and estimate its intensity. The resonance of electric circuits is illustrated in the tuned circuits of the radio, which respond only to sending stations of a particular wave length, and practically not at all to other stations. In optics, the theory of refractive index and absorption coefficient is closely connected with resonance. As is shown by the sharp spectrum lines, atoms contain oscillators capable of damped simple harmonic motion, or at any rate act as if they did; the real theory, using wave mechanics, is complicated but leads essentially to this result. An external light wave is a sinusoidal impressed force, leading to a forced motion of the oscillators with the same frequency but different phase. The component of motion in phase with the field reacts back on the field to change its phase, and this progressive change of phase as the light travels through the body is interpreted as a changed velocity of propagation, or an index of refraction different from unity. Similarly the other component produces a diminution of intensity, or absorption. The phenomenon of anomalous dispersion, with abnormally large index of refraction and absorption coefficient, comes about when the external wave is in resonance with the atom.

22. Mechanical Resonance.—Let the external force be $F_0 \cos \omega t$. It is simpler to regard this as being the real part of $F_0 e^{i\omega t}$. Thus we use the differential equation

$$m \frac{d^2 x}{dt^2} + 2mk \frac{dx}{dt} + m\omega_0^2 x = F_0 e^{i\omega t}, \quad (7)$$

where we use ω_0 for the natural angular frequency, to distinguish from the impressed angular velocity ω . The resulting x will be complex, and its real part represents the actual motion. Now we assume that the forced motion has the same frequency as the impressed force, or that $x = Ae^{i\omega t}$, where A may be complex. If A_r and A_i are the real and imaginary parts of A , we easily see that the real part of x is given by

$$A_r \cos \omega t - A_i \sin \omega t, \quad (8)$$

so that in general the motion has one term in phase with the force, whose amplitude is given by the real part of A , and another out of phase, the amplitude being the negative of the imaginary part. Substituting our exponential formula for x in Eq. (7), we have

$$[m(-\omega^2) + 2mk(i\omega) + m\omega_0^2]Ae^{i\omega t} = F_0e^{i\omega t}.$$

Canceling the exponential, we have

$$A = \frac{F_0}{m} \frac{1}{(\omega_0^2 - \omega^2) + 2ik\omega}. \quad (9)$$

To get the coefficients of terms in phase and out of phase with the force, or A_r and $-A_i$, we multiply numerator and denominator by the conjugate of the denominator, obtaining respectively

$$A_r = \frac{F_0}{m} \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + 4k^2\omega^2}$$

and

$$-A_i = \frac{F_0}{m} \frac{2k\omega}{(\omega_0^2 - \omega^2)^2 + 4k^2\omega^2}. \quad (10)$$

These two functions are plotted in Fig. 4. It is seen that the first has the form made familiar by the anomalous dispersion curve in optics, the second resembling the corresponding absorption curve. This resemblance is an essential one, as we shall see in Chap. XXIV. One feature of the curves should be mentioned. The anomalous behavior in the neighborhood of ω_0 is confined to a narrower and narrower band of frequencies as k becomes smaller and smaller compared with ω_0 , so that if the damping is very small the resonance is very sharp, while if there is large damping, there is a broad range of frequencies over which resonance is appreciable.

23. Electrical Resonance.—Suppose that a dynamo supplies sinusoidally alternating electromotive force, given by $E \cos \omega t$,

to an electric circuit containing resistance, inductance, and capacity. The differential equation for the charge is then

$$L \frac{d^2 q}{dt^2} + R \frac{dq}{dt} + \frac{q}{C} = E \cos \omega t. \quad (11)$$

We set up instead the differential equation for the current $i =$

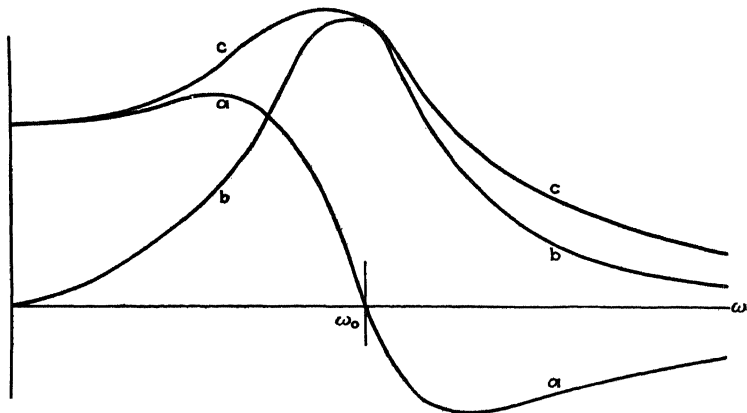


FIG. 4.—Amplitude of forced motion of an oscillator, as function of frequency. (a) Component in phase with force; (b) component out of phase.

dq/dt , which we obtain from Eq. (11) by differentiating with respect to time:

$$L \frac{d^2 i}{dt^2} + R \frac{di}{dt} + \frac{i}{C} = \frac{d}{dt}(E \cos \omega t). \quad (12)$$

As with the mechanical case, we replace $E \cos \omega t$ by the complex exponential $E e^{i\omega t}$, of which the real part gives the electromotive force. Similarly we assume the current to be sinusoidal, given by the real part of $i_0 e^{i\omega t}$. Making these changes in Eq. (12), and carrying out the differentiations, we have

$$\left(L \frac{d^2}{dt^2} + R \frac{d}{dt} + \frac{1}{C} \right) i_0 e^{i\omega t} = i\omega E e^{i\omega t},$$

$$i_0 = \frac{E}{R + i(L\omega - 1/C\omega)}. \quad (13)$$

The denominator here equals $Z e^{i\alpha}$, where

$$Z = \sqrt{R^2 + X^2}, \quad X = L\omega - \frac{1}{C\omega}, \quad (14)$$

and

$$\alpha = \tan^{-1} \frac{X}{R},$$

where X is called the reactance, Z the impedance. Then the current is

$$i = \frac{E}{Z} \cos (\omega t - \alpha). \quad (15)$$

The impedance takes the place of the resistance in problems involving alternating currents, since we divide the amplitude of the e.m.f. by the impedance rather than by the resistance to get the amplitude of the current. We note that the impedance is a function of frequency. It becomes infinite when the frequency becomes zero, on account of the term involving the capacity, and showing that a direct current cannot go through a condenser; and also when the frequency becomes infinite, on account of the term in the inductance, showing that infinitely rapid oscillations cannot pass through the inductance. In between, it goes through a minimum, at the frequency for which $X = 0$, or $\omega = 1/\sqrt{LC}$, the natural frequency at which the circuit would oscillate by itself if it had no resistance or impressed e.m.f. Thus for impressed e.m.fs. of the same amplitude, but of a variety of frequencies, that whose frequency agrees most closely with the natural frequency will produce the largest current, and the others may produce much smaller currents, so that we have resonance, or tuning. To tune a circuit, one adjusts L or C , or both. When it is tuned, the sharpness of tuning depends on the size of R . For instance, if R were 0, there would be infinite response at exact resonance, so that the tuning would be infinitely sharp.

In addition to the dependence of amplitude on frequency, there is also a phase difference between e.m.f. and current, given by the quantity α above. We can get a simple interpretation of this in the complex plane. The quantity $R + iX$ is called the complex impedance. Its magnitude is just the real impedance Z , and its phase, or angle, is the angle α . It is interesting to note that α goes from -90 deg. at zero frequency to $+90$ deg. at infinite frequency, passing through zero at resonance.

24. Superposition of Transient and Forced Motion.—The general solution of an oscillatory problem is the sum of the steady-state motion (the particular solution); and a transient with arbitrary amplitude and phase, chosen to satisfy the initial conditions. Thus, choosing an electrical case, we may have no charge and current in a circuit at $t = 0$, but start applying a

sinusoidal e.m.f. at that instant. The charge and current at any later time are given by

$$q = A e^{-\frac{R}{2L}t} \cos(\omega_0 t - \alpha_0) + \frac{E}{\omega Z} \sin(\omega t - \alpha),$$

$$i = -A \omega_0 e^{-\frac{R}{2L}t} \sin(\omega_0 t - \alpha_0) - A \frac{R}{2L} e^{-\frac{R}{2L}t} \cos(\omega_0 t - \alpha_0) + \frac{E}{Z} \cos(\omega t - \alpha),$$

where ω_0 is the natural angular frequency, A and α_0 the amplitude and phase of the transient. Then to determine A and α_0 we have the equations

$$0 = q_0 = A \cos \alpha_0 - \frac{E}{\omega Z} \sin \alpha$$

$$0 = i_0 = A \omega_0 \sin \alpha_0 - A \frac{R}{2L} \cos \alpha_0 + \frac{E}{Z} \cos \alpha, \quad (16)$$

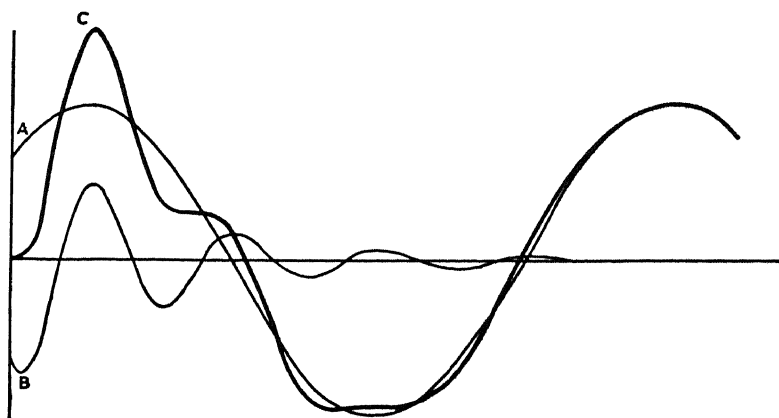
where q_0, i_0 are initial charge and current, equal to zero for these particular initial conditions.

Three examples of the charge as a function of time are given in Fig. 5. In (a), the natural frequency is taken to be much greater than the external frequency, and the logarithmic decrement large, so that the transient is a rapidly damped high frequency vibration, which is imperceptible after a few periods of the external force. The case (b) is that in which external and natural frequencies are almost equal, and the damping small. In this case, the forced and transient vibrations, having almost the same frequencies, form beats with each other, as one always has when two almost equal frequencies are superposed, the sum of two sine waves leading to a sinusoidal vibration whose frequency is the average of the two frequencies, but whose amplitude is modulated with the slow difference frequency between the two vibrations, as given by the equation

$$\cos \omega_1 t + \cos \omega_2 t = 2 \cos \left(\frac{\omega_1 - \omega_2}{2} \right) t \cos \left(\frac{\omega_1 + \omega_2}{2} \right) t. \quad (17)$$

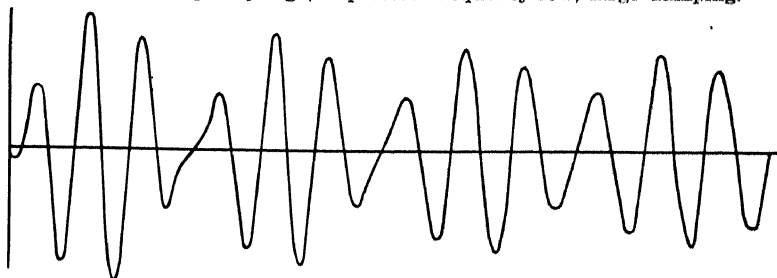
Since the transient gradually dies down, however, the amplitude of the beats grows less and less, until gradually only the forced motion remains. In the case (c), the external frequency is exactly equal to the natural frequency. Here there are no beats,

the amplitude merely building up exponentially to its final value.

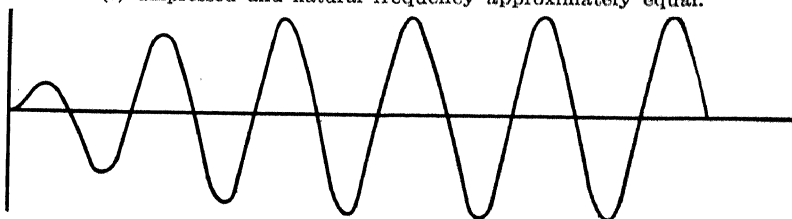


Curve *A* is forced motion, *B* transient, *C* combined motion.

(a) Natural frequency high, impressed frequency low, large damping.



(b) Impressed and natural frequency approximately equal.



(c) Impressed and natural frequency equal.

FIG. 5.—Transient and forced motion superposed.

25. Motion under General External Forces.—If we are given an arbitrary external force, say $F(t)$, we shall show in a later chapter that it is possible to write it as a sum of sinusoidal terms:

$$F(t) = \text{real part of } \sum_n F_n e^{i\omega_n t}.$$

Thus any sound may be considered as made up of a superposition of pure tones, and any light as a superposition of pure colors. Now suppose we find the forced motion resulting from each of these sinusoidal vibrations acting separately, and then add them. The result will be the solution of the whole problem. For suppose $x_n(t)$ is the solution of the problem whose force is the n th term of the summation, so that we have

$$\left(m \frac{d^2}{dt^2} + 2mk \frac{d}{dt} + m\omega_0^2\right)x_n = F_n e^{i\omega_n t}.$$

Add all these equations. Then we have

$$\left(m \frac{d^2}{dt^2} + 2mk \frac{d}{dt} + m\omega_0^2\right) \sum_n x_n = \sum_n F_n e^{i\omega_n t},$$

showing that $\sum_n x_n$ satisfies the whole equation. We readily

see that this is a special case of a general theorem: if the impressed force, in an inhomogeneous linear equation, is written as a sum of terms, and if we have solutions of the separate problems in which only one term of the sum is impressed at a time, the solution of the whole problem is the sum of these separate solutions. We note that those particular forces whose frequencies are near the natural frequency will produce greatly exaggerated responses.

26. Generalizations Regarding Linear Differential Equations.

We have made several generalizations regarding linear differential equations, and it is well to group these together. We have seen that

1. Any linear combination of solutions of a homogeneous linear differential equation is itself a solution, and if the linear combination contains as many arbitrary constants as the order of the differential equation, it is a general solution.

2. A general solution of an inhomogeneous linear differential equation is the sum of a particular solution, and a general solution of the corresponding homogeneous equation.

3. If the inhomogeneous part of an inhomogeneous linear differential equation is a sum of terms, and if we have the solutions of the equations formed by taking just one of these separately, the particular solution of the whole problem can be formed by adding these separate solutions.

Physically, the first statement means that free vibrations of a system governed by a linear differential equation may be super-

posed without affecting each other. The second means that free vibrations can coexist with forced vibrations; and the last, that forced vibrations from different sources can coexist without affecting each other. All these properties of coexistence or superposability of vibrations are characteristic only of linear equations, but, as we shall see, a great many physical phenomena are governed by such equations, so that the superposability of vibrations is of widespread physical importance.

Problems

1. A coil of resistance 2 ohms, inductance 10 millihenries, is connected to a condenser of capacity 10 mf. At $t = 0$, the condenser is charged to a potential of 100 volts, and no current is flowing. Find the charge on the condenser at any later time, and also the current flowing. What are the period and logarithmic decrement of the circuit? What would the resistance have to be, leaving inductance and capacity the same, such that the system would be critically damped?

2. Prove that the displacement of a particle in damped oscillation is given by

$$x = e^{-kt} \left(x_0 \cos \sqrt{\omega^2 - k^2} t + \frac{v_0 + kx_0}{\sqrt{\omega^2 - k^2}} \sin \sqrt{\omega^2 - k^2} t \right),$$

where x_0 , v_0 are initial values of displacement and velocity. Pass to the case of critical damping, by letting $\omega^2 - k^2$ approach zero. Show that the resulting motion has one term of the form te^{-kt} , and prove directly that this satisfies the differential equation.

3. Letting $\omega = k/2$, draw curves for x as a function of t , representing the damped motion for the case where the initial velocity is zero but the initial displacement is not, and also for the case where the initial displacement is zero but the velocity is not.

4. A pendulum is damped so that its amplitude falls to half its value in 1 min. Its actual period is 2 sec. Find the change in the period which there would be if the damping were not present. (Hint: use power series expansion for frequency, treating k as a small quantity.)

5. A radio receiving station has a circuit tuned to a wave length of 500 m. It is desired to have the tuning sharp enough so that a frequency differing from this by 10,000 cycles per second gives only 1 per cent as much response as the natural frequency, for the same amplitude of signal. Work out reasonable values of resistance, inductance, and capacity to accomplish this.

6. The sharpness of tuning of a vibrating system may be measured by the so-called half breadth of the resonance band, or the frequency difference between the two frequencies for which the amplitude of response is half that at exact resonance. Prove that the ratio of half breadth to resonance frequency is proportional to the logarithmic decrement, if the damping is not too great.

7. A tuning fork of pitch C (256 vibrations per second) is so slightly damped that its amplitude after 10 sec. is 10 per cent of the original amplitude. It is set into oscillation, first by another fork of the same pitch, then

by one a semitone higher, both vibrating with the same amplitude. Find the ratio of amplitudes of forced motion in the two cases. What will be the pitch of the forced vibration in the second case?

8. The support of a simple pendulum moves horizontally back and forth with simple harmonic motion. Show that this sets the pendulum into forced motion, as if there were a force applied directly to the bob. Show that the motion has the following behavior: The pendulum pivots about a point not its point of support, but such that, if it were really pivoted here, its natural period would be the actual period of the forced motion. Discuss the cases where the pivotal point is below the point of support; above the point of support. Neglect transients.

9. A particle subject to a linear restoring force and a viscous damping is acted on by a periodic force whose frequency differs from the natural frequency by a small quantity. The particle starts from rest at $t = 0$, and builds up the motion. Discuss the whole problem, including initial conditions. Consider what happens in the limiting case when the frequency gets nearer and nearer the natural frequency, and the damping gets smaller and smaller. Show that the results are as indicated in Fig. 5, (b), (c).

10. The amplitude of the forced current in a circuit is

$$i_0 = \frac{E}{[R + i(L\omega - 1/C\omega)]}$$

Plot real part as abscissa, imaginary part as ordinate, obtaining a curve by taking points for all frequencies. Find the equation of the resulting curve, and prove that it is a circle.

11. Show that for a particle subject to a linear restoring force and viscous damping the maximum amplitude occurs when the applied frequency is less than the natural frequency. Find this resonance frequency. Show that maximum energy is attained when the applied frequency equals the natural frequency. What are the maximum amplitude and maximum energy?

12. The motion of an anharmonic undamped oscillator is described by

$$m \frac{d^2x}{dt^2} + m\omega_0^2 x + bx^2 = 0,$$

where b is a small quantity. Solve this equation by successive approximations, expanding x in a power series in powers of b .

13. If the oscillator in Problem 12 is acted on by a force $A \cos pt + B \cos qt$, show that the steady-state solution contains terms of frequencies $2p$, $2q$, $q + p$, $q - p$, $2q + p$, $2q - p$, etc. Note that superposition does not hold for the equation above. These new frequencies are called combination tones.

CHAPTER V

ENERGY

We have progressed far enough in our study of mechanics so that it will pay to stop and survey the situation. Mechanics is a large subject, and we may consider some of the directions in which we could extend what we have done already. In the first place, we may treat the mechanics of many sorts of systems. We may have the mechanics of particles, or of rigid bodies, or of deformable, elastic solids, or of fluid media. All these we shall treat, in more or less detail, before we are through. What we have done so far comes under the heading of mechanics of particles, and we shall look at that field in more detail.

In the first place, one almost never has real particles to deal with in a mechanical problem. Probably the closest approach is found in the kinetic theory of monatomic gases, where the atoms act like movable points exerting forces on each other. But often very large bodies can act as particles, as, for instance, the planets in their motions about the sun. Then again we can have essentially complicated systems, like pendulums, or weights suspended on springs, which yet have such simple motions that we can apply the methods of the mechanics of particles to them. Many of the problems we have treated so far have been of this sort.

A particle has three coordinates, which may be x , y , z , and the problem of mechanics is to find the way in which these coordinates change with time. The starting point is Newton's second law of motion, giving the accelerations, or second time derivatives of the coordinates, in terms of the forces. All of our problems so far, whether dealing with actual particles or not, fall under this classification, and in fact belong to the more restricted class of one-dimensional problems, with but one coordinate x . The next few chapters will be devoted to the two- and three-dimensional cases of mechanics of a particle.

The one-dimensional motions of a particle fall into different classes, depending on the type of force acting. We have treated several sorts of forces: viscous resistances, linear restoring forces,

external forces which are arbitrary functions of time. That is, the force may be a function of velocity, of position, or of time, or, of course, of all three combined. Most common mechanical problems are of this type, the force depending on v , x , and t , but this is not necessary. For instance, in radiation problems, in electromagnetic theory, one meets a force proportional to the time derivative of acceleration, or to d^2x/dt^2 , which turns out to act much like a viscous resistance. But such cases are rare.

The simplest cases are those in which the force depends only on the coordinate. Then, in one-dimensional motion, we can always introduce a potential energy, which added to the kinetic energy gives a total energy that stays constant, expressing the conservation of energy. If, on the other hand, there are external impressed forces, the energy may increase or decrease with time, depending on whether the impressed forces do work on the system or have work done on them; while, if there are frictional forces, the energy will decrease with time, being dissipated in heat, for which reason these forces are called dissipative forces. It is plain that the study of different types of forces is closely tied up with the idea of energy, which we so far have not discussed, and we turn to this question, first deriving the mathematical formulation of kinetic energy for one-dimensional problems.

27. Mechanical Energy.—Let us see where the concept of energy comes from, and how we can use it. We start with a particle of mass m , acted on by a force F . Then Newton's second law is $m d^2x/dt^2 = F$. Now let us multiply each side by dx/dt , and integrate with respect to t , from time t_0 up to t :

$$m \int_{t_0}^t \frac{dx}{dt} \frac{d^2x}{dt^2} dt = \int_{t_0}^t F \frac{dx}{dt} dt.$$

Both these integrals can be transformed. First, we note that

$$\frac{d}{dt} \left(\frac{dx}{dt} \right)^2 = 2 \frac{dx}{dt} \frac{d^2x}{dt^2}.$$

Thus the left side is

$$\frac{m}{2} \int_{t_0}^t \frac{d}{dt} \left(\frac{dx}{dt} \right)^2 dt = \frac{m}{2} \left(\frac{dx}{dt} \right)^2 \Big|_{t_0}^t,$$

or letting dx/dt be denoted by v , and its value at $t = t_0$ by v_0 , this side is $mv^2/2 - mv_0^2/2$. On the right, $\int F dx/dt dt = \int F dx$, where

now the integral is from x_0 to x , if x_0 is the value of x at $t = t_0$, x at t . Then the equation is

$$\frac{1}{2}mv^2 - \frac{1}{2}mv_0^2 = \int_{x_0}^x F dx. \quad (1)$$

The quantity $mv^2/2$ is called the kinetic energy, $\int F dx$ is the work done, and our equation says that the work done by the force on the particle between two instants of time equals the increase in kinetic energy during the time. This is the fundamental proposition relating to energy, and our proof is the standard one.

Next we consider the nature of the force F . First there is the case where it depends only on the position of the particle, as in a gravitational field or a linear restoring force, without friction. Then $F = F(x)$, and we may write $\int_{x_0}^x F(x) dx = -V(x)$, so that $mv^2/2 + V(x) = mv_0^2/2 + V(x_0)$. The quantity $V(x)$ is called the potential energy, and the sum of it and the kinetic energy is the total energy; our equation states that the total energy remains constant during the motion. The lower limit x_0 of integration may be chosen in an arbitrary way, or an arbitrary constant of integration may be added to the potential energy, without changing the results, which depend only on potential differences. The potential energy is related to the force either by the equation above, or by its derivative, $F = -dV/dx$.

In case the force depends on the velocity as well as the position, the situation is quite different. Then the value of F cannot be predicted when x is known, so that we cannot even evaluate the work done without knowing more details about the system. In such a case it is plainly impossible to set up a potential energy function independent of time, or to speak of the total energy being conserved. Such a system is called nonconservative, in contrast to a conservative system in which the energy stays constant. Even in a nonconservative system it is often possible to write a potential function connected with part of the force. Thus with a damped oscillator, we can write a potential function for the restoring force, but not for the viscous resistance. In such a case we shall still speak of the sum of the kinetic and potential energy as being the total energy, but we can no longer say that it remains constant. Rather we should say that the time rate of change of the energy was equal to the rate of working

of outside forces, both of viscosity and of any external impressed forces, on the system. Let us see what this means mathematically. Let $F = -dV/dx + G$, where V is the potential function for that part of the force derivable from a potential, and G is the remaining force. Then the energy is $mv^2/2 + V$. The time rate of change of the energy is

$$\frac{1}{2}m\left(2v\frac{dv}{dt}\right) + \frac{d}{dt}[V(x)] = \left(m\frac{d^2x}{dt^2} + \frac{dV}{dx}\right)v,$$

using $\frac{dV(x)}{dt} = \frac{dV(x)}{dx} \frac{dx}{dt}$. But $\left(m\frac{d^2x}{dt^2} + \frac{dV}{dx}\right) = G$, by Newton's second law, so that the time derivative of energy reduces to Gv , or the external force times the velocity, as we should expect.

One should not be disturbed to find systems whose total energy does not stay constant. At first sight they seem to contradict the general law of conservation of energy, but on closer examination we always find that they are parts of a larger system whose energy really is conserved. Thus if we consider not merely the damped vibrating particle but also the viscous fluid doing the damping, we shall find that the latter gains the energy lost by the former, transforming it into heat, itself a form of energy. It is in fact a general situation that there are two ways of treating a mechanical problem: first, by considering the whole system, and treating it as a conservative system; second, treating only part of the system, and taking the forces exerted by the rest on this part as being impressed or dissipative forces, which cannot be derived from a potential.

28. Use of the Potential for Discussing the Motion of a System.

The one-dimensional motion of a particle in a conservative field can be discussed with great ease by the use of the potential function. Suppose we know V as a function of x , and suppose that we inquire about the motion of a particle of total energy E in this potential field. Then we have $mv^2/2 = E - V$, $v = \sqrt{2(E - V)}/m$. Since this is a known function of x , we can find the speed at every point. In the first place, we can use this to get an explicit solution of the problem. For writing $v = dx/dt$, and integrating, we have

$$t - t_0 = \int \frac{dx}{\sqrt{2(E - V)/m}}, \quad (2)$$

giving a relation between t and x , involving two arbitrary con-

stants (energy, and the constant of integration, determining the origin of time, or the phase). Thus for instance for a particle moving under gravity, where $V = mgx$, we have

$$t - t_0 = \int \frac{dx}{\sqrt{2E/m - 2gx}}.$$

Letting $z = 2E/m - 2gx$, so that $dz = -2gdx$, this is

$$t - t_0 = -\frac{1}{2g} \int \frac{dz}{\sqrt{z}} = -\frac{\sqrt{z}}{g} = -\frac{1}{g} \sqrt{2E/m - 2gx},$$

where evidently t_0 is the value of t at which $2E/m - 2gx = 0$, or $x = E/mg$, which, as we readily see, is the highest point of the path, at which the body commences to fall. If we let this value of x be x_0 , then, squaring, we have $x - x_0 = -\frac{1}{2}g(t - t_0)^2$, the familiar solution. Many one-dimensional problems can be solved by this method, as for instance the pendulum with large amplitude, which leads to an elliptic integral. On the other hand, there are, of course, many cases where the integration is too difficult to carry out.

Even if the solutions cannot be obtained exactly, however, we can still use the method of the energy to get general information about the problem. Let us imagine V plotted as a function of x (see Fig. 6). Then we draw on the same graph a horizontal line at height E . The square root of the difference between the two curves is then proportional to the velocity of the particle at that point. Thus the velocity is only real where this difference is positive, and is imaginary elsewhere. If the velocity is only real in certain regions of x , this means that the motion can only occur within those regions. As the particle approaches the edge of such a region, the speed gets smaller and smaller, and finally at the edge the particle stops. Then it reverses and travels away again. The possibility of going either toward or away from the boundary comes from the two signs of the square root: the velocity at a given point of space is always the same in magnitude, but can be in either direction. If now the region where the kinetic energy is positive is bounded at both ends, then after reversing its motion at one edge, the particle will travel to the other, reverse, come back, and repeat the process indefinitely. Since at a given point the particle always travels with the same speed, it will always require the same time to traverse its path, and the motion will be periodic. Thus, if the total

energy is E_1 (Fig. 6), the motion is periodic, confined between c and e . If it is E_2 , either of two periodic motions is possible, between b and f , or between h and j . This is a general result for a conservative motion in one dimension which does not extend to infinity.

If, on the other hand, the kinetic energy remains positive in one direction all the way to infinity, but becomes negative at a finite point in the other direction, the particle will come in from

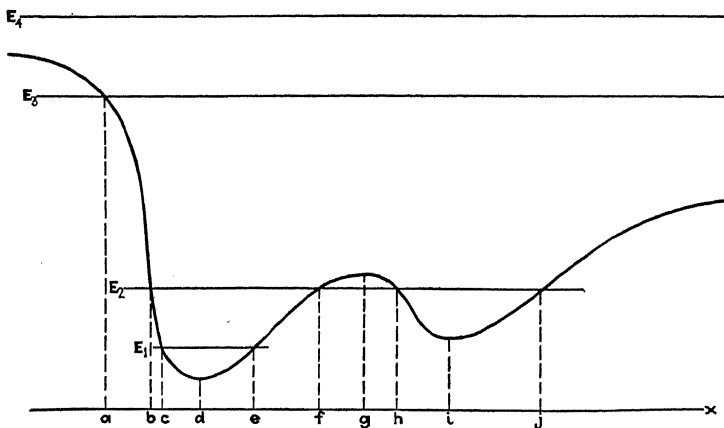


FIG. 6.—Potential energy V as function of coordinate x .

Total energy E_1 , periodic motion between c and e .

Total energy E_2 , periodic motion between b and f , or h and j .

Total energy E_3 , nonperiodic motion between a and infinity, reversing at a .

Total energy E_4 , nonperiodic nonreversing motion.

infinity, having taken since negatively infinite time to do it, will reverse, and will return to infinity. This is the case with energy E_3 , the particle coming in from the right, speeding up in the region about i , slowing down about g , speeding up about d , finally coming to rest at a , and reversing, traveling back to the right. An example of the first, periodic case is a particle vibrating in simple harmonic motion, and of the second nonperiodic case is a ball coming from infinity, hitting a wall, and being bounced back again, or a ball thrown up in the air and coming down again. Finally there is the possibility of a potential such that the kinetic energy is positive at all values of x . Then the particle persists in one direction forever, like a free particle, but generally travels with a variable velocity. Such a case is found for energy E_4 , where the particle starts from infinite distance in one direction, travels toward the center, speeds up and slows

down corresponding to the maxima and minima, and finally goes to infinity in the other direction. It is to be noted that motions in the same potential field, but with different total energy, can have quite different characteristics under this classification. Thus oscillatory motions are always possible around minima in the potential energy, for small enough total energy. But it may be that for too high total energy the particle will be able to get entirely away from the neighborhood of the minimum, and will go to infinity. In Fig. 6, there are three points, d , g , and i , at which the force is zero, and the particle would stay at rest forever, if it were placed at one of these points. Of these, g is a position of unstable equilibrium, and a small impact would start the particle oscillating, about either d or i . On the other hand, d and i are both points of stable equilibrium, so that a particle at rest at either of these points would suffer only small oscillations about that point if struck a small impact.

29. The Rolling-ball Analogy.—A simple model which shows the properties of one-dimensional motion can be set up as follows. We imagine a track, like a roller coaster, set up, shaped just like the potential curve. Then we start a ball rolling on this track, starting from rest at a given height. Its motion will then approximate that of a particle in the corresponding potential field. The reason is that, since gravitational potential is proportional to height, the ball actually has the potential at any point which it should, and correspondingly the correct speed. The only approximations made, other than friction, consist in neglecting the fact that part of the kinetic energy actually goes into up and down motion, and part into rotation, instead of all into horizontal motion. From such a model, we can see how motion may be oscillatory, if the track rises on the far side of a dip up to the height where the ball started, or how it can go to infinity if the track continues permanently at a lower level. We can also see the general character of the solution in the case where there is damping, just by imagining that the ball is subject to friction. Obviously the motion still will have the character of the undamped motion, but corresponding to a continually decreasing energy. Thus with an oscillatory motion the amplitude will constantly decrease until it stops, while with a motion which originally was not oscillatory it may be possible that it become trapped in a minimum of potential, settle down to oscillate, and eventually come to rest. In any case, if the damping

continues, the motion will eventually stop, at a minimum of potential.

30. Motion in Several Dimensions.—So far, we have treated only the motion of a particle in one dimension. If it can move about in two- or three-dimensional space, however, the problem becomes much more difficult. Suppose the coordinates of a particle are x, y, z , so that its motion is described by finding x, y , and z as functions of time. Force, acceleration, are vectors, and our first task is to investigate vector analysis enough to deal with these quantities. We shall find that in two and three dimensions it is by no means true that all force fields, in which the force is a function of the position alone, can be derived from a potential function. The next chapters, then, will deal with vectors, force fields, and potentials. When we come to the equations of motion, we find separate equations for each component: if F_x, F_y, F_z represent the components of force along the axes, we have

$$m \frac{d^2x}{dt^2} = F_x, m \frac{d^2y}{dt^2} = F_y, m \frac{d^2z}{dt^2} = F_z, \quad (3)$$

a set of simultaneous differential equations. Such equations can be solved in a few simple cases. For instance, if F_x depends only on x , F_y only on y , F_z only on z , they are simply three independent equations, which we can handle by the methods already used. This is called the method of separation of variables, and much of our effort will be directed toward this method of solution. We shall carry out methods of changing to arbitrary coordinate systems, with a view to separating variables. For instance, in motion under a force acting toward a center of attraction, we introduce polar coordinates, and in these the equation for r is separated from those for the angles, so that we can solve. The process of changing coordinate systems leads us to Lagrange's equations, the equations of motion in generalized coordinates. Finally, the method of energy, the rolling-ball analogy, and the other methods of the present chapter, can be used in several dimensions, and provide the best means for a qualitative discussion of a problem.

Problems

1. Take the sinusoidal solution for the displacement of a harmonic oscillator, find the velocity from it, compute kinetic and potential energy as func-

tions of time, and add to show that the sum remains constant. Show that the energy is proportional to the square of the amplitude.

2. Proceed as in Prob. 1, but for the damped oscillator, finding the sum of kinetic and potential energies, showing that it decreases with time. Compute the time rate of change of the energy, find the rate of working of the frictional force, and show by direct comparison that they are equal.

3. Let a particle move in a field whose potential is $-1/x + 1/x^2$. Show by graphical methods that for small total energy the motion is oscillatory, but that for larger energy it is nonperiodic and extends to infinity. Find the energy which forms the dividing line between these two cases. Compute the limiting frequency of the oscillatory motion as the amplitude gets smaller and smaller (using the results of Prob. 1, Chap. I), and describe qualitatively how the frequency changes when the amplitude increases.

4. Solve directly the problem of the motion of a particle moving in a field of potential $-1/x + 1/x^2$, using the energy integral. Show that the mathematical solution has the physical properties found in Prob. 3.

5. Using the solution of Prob. 4, find the period of oscillation of the oscillatory solutions in the potential $-1/x + 1/x^2$, as functions of the energy. To do this, note that the two ends of the path are the values of x for which

$\sqrt{2(E - V)/m} = 0$. Thus the integral $\int_{x_0}^{x_1} \frac{dx}{\sqrt{2(E - V)/m}}$ from one of these points to the other will give just the half period. Show that the period approaches the value found in Prob. 3 for small oscillations.

6. In an electric circuit, show that one can set up a magnetic energy $\frac{1}{2}Li^2$ analogous to a kinetic energy, and an electric energy $\frac{1}{2}q^2/C$ analogous to a potential energy. Show that the rate of change of this total energy equals the rate of working of the resistance and the applied electromotive force.

7. An atom acts like a particle held to a position of equilibrium by a definite restoring force and a viscous resistance. An external light wave exerts a sinusoidal force, the atom executing a forced vibration under the influence of the wave. Show that the atom continually absorbs energy from the wave, the energy going into the viscous resistance. Show that the rate of absorption is proportional to the component of amplitude out of phase with the force, which we have already connected with the absorption coefficient.

8. Solve the problem of the undamped oscillator, by using the equation $t = \int dx / \sqrt{2(E - V)/m}$.

9. Discuss the problem of the pendulum with arbitrary amplitude by the graphical method. Show that for low energies the motion is oscillatory, but for high energies it is a continuous rotation. Sketch the qualitative form of curves for angular displacement as a function of time, for several energies, in both the oscillatory and rotatory ranges.

10. Set up the problem of the pendulum by the method of Prob. 8, and show that t as a function of the angle is given by an elliptic integral. (Hint: Use the information about elliptic integrals given in Peirce's table; note that $1 - \cos \theta = 2 \sin^2 \frac{1}{2}\theta$.)

CHAPTER VI

VECTOR FORCES AND POTENTIALS

In our one-dimensional problems, we have had no occasion to mention vectors; however, before we can treat the detailed theory of motion in two or three dimensions, we must discuss them, and their relation to such things as potential energy.

31. Vectors and Their Components.—The force, in two- or three-dimensional motion, is a vector, and we must make a study of the mathematical relations of vectors. In the first place, a vector is often denoted by its components along three axes at right angles, as F_x, F_y, F_z . Vectors, in the second place, obey the following law of addition: if two vectors F and G have components F_x, F_y, F_z and G_x, G_y, G_z , respectively, the components of the sum $F + G$ are $(F_x + G_x), (F_y + G_y), (F_z + G_z)$. A graphical discussion shows that this is equivalent to the familiar parallelogram law of addition (as in Fig. 2, where the same proposition was shown for complex numbers, regarded as vectors). Third, if we multiply a vector by a constant, as C , each component is multiplied by this constant. Thus the components of CF are CF_x, CF_y, CF_z . Often a constant like C is called a "scalar," to distinguish it from a vector. A scalar is a quantity which has magnitude but not direction, a vector having both magnitude and direction.

It is often useful to write vectors in terms of three so-called unit vectors, i, j, k . Here, i is a vector of unit length, pointing along the x axis, and similarly j has unit length and points along the y axis, and k along the z axis. Now we can build up a vector F out of them, by forming the quantity $iF_x + jF_y + kF_z$. This is the sum of three vectors, one along each of the three axes, and the first, which is just the component of the whole vector along the x axis, is F_x , and the other components likewise are F_y and F_z . Thus the final vector has the components F_x, F_y, F_z , and is just the vector F .

By the magnitude of a vector we mean its length. By the three-dimensional analogy to the Pythagorean theorem, by which the square on the diagonal of a rectangular prism is the sum

of the squares on the three sides, the magnitude of a vector F equals $\sqrt{F_x^2 + F_y^2 + F_z^2}$. We often speak of unit vectors, *i.e.*, vectors whose magnitude is 1.

The component of a vector in a given direction is simply the projection of the vector along a line in that direction. It evidently equals the magnitude of the vector, times the cosine of the angle between the direction of the vector and the desired direction. As a special example, the component of a vector F along the x axis is F_x , and this must equal the magnitude of F , times the cosine of the angle between F and x . If this angle is called (F, x) , then we must have $\cos (F, x) = \frac{F_x}{\sqrt{F_x^2 + F_y^2 + F_z^2}}$,

with similar formulas for y and z components. The three cosines of the angles between a given direction, as the direction of the vector F , and the three axes, are called direction cosines, and are often denoted by letters l, m, n , so that in this case we have $l = \cos (F, x)$, etc. It follows immediately that $l^2 + m^2 + n^2 = 1$. We can make a simple interpretation of the direction cosines of any direction: they are the components of a unit vector in the desired direction, along the three coordinate axes.

32. Scalar Product of Two Vectors.—Multiplication of two vectors is a rather special process, and there are two entirely independent products, called the “scalar product” and the “vector product.” We shall first consider the scalar product. The scalar product of two vectors F and G is denoted by $F \cdot G$, and by definition it is a scalar, equal to either (1) the magnitude of F times the magnitude of G times the cosine of the angle between; or (2) the magnitude of F times the projection of G on F ; or (3) the magnitude of G times the projection of F on G . From the last section we see that these definitions are equivalent. It is often useful to have the scalar product of two vectors in terms of the components along x, y , and z . We find this by writing in terms of i, j , and k . Thus we have

$$\begin{aligned} F \cdot G &= (iF_x + jF_y + kF_z) \cdot (iG_x + jG_y + kG_z) \\ &= (i \cdot i)F_xG_x + (i \cdot j)F_xG_y + (i \cdot k)F_xG_z \\ &\quad + (j \cdot i)F_yG_x + (j \cdot j)F_yG_y + (j \cdot k)F_yG_z \\ &\quad + (k \cdot i)F_zG_x + (k \cdot j)F_zG_y + (k \cdot k)F_zG_z. \end{aligned}$$

But now by the fundamental definition,

$$\begin{aligned} i \cdot i &= j \cdot j = k \cdot k = 1, \\ i \cdot j &= j \cdot i = j \cdot k = k \cdot j = k \cdot i = i \cdot k = 0. \end{aligned}$$

Thus

$$F \cdot G = F_x G_x + F_y G_y + F_z G_z. \quad (1)$$

The scalar product has many uses, principally in cases where we are interested in the projections of vectors. For example, the scalar product of a vector with a unit vector in a given direction equals the projection of the vector in the desired direction. The scalar product of a vector with itself equals the square of its magnitude, and is often denoted by F^2 . The scalar product of two unit vectors gives the cosine of the angle between the directions of the two vectors. To prove that two vectors are

at right angles, we need merely prove that their scalar product vanishes.

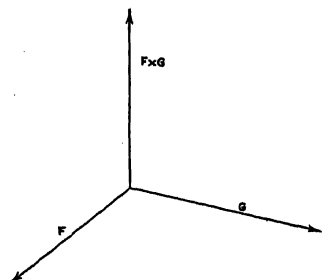


FIG. 7.—Direction of the vector product.

33. Vector Product of Two Vectors.

The vector product of two vectors F and G is denoted by $(F \times G)$, and by definition it is a vector, at right angles to the plane of the two vectors, equal in magnitude to either (1) the magnitude of F times the magnitude of G times the sine of the angle between them; or (2) the mag-

nitude of F times the projection of G on the plane normal to F ; or (3) the magnitude of G times the projection of F on the plane normal to G . We must further specify the sense of the vector, whether it points up or down from the plane. This is shown in Fig. 7, where we see that F , G , and $F \times G$ have the same relations as the coordinates x , y , z in a right-handed system of coordinates. Another way to describe the rule in words is that, if one rotates F into G , the rotation is such that a right-handed screw turning in that direction would be driven along the direction of the vector product. From this rule, we note one interesting fact: if we interchange the order of the factors, we reverse the vector. Thus $(F \times G) = -(G \times F)$.

We can compute the vector product in terms of the components, much as we did with the scalar product. Thus we have

$$\begin{aligned} F \times G &= (iF_x + jF_y + kF_z) \times (iG_x + jG_y + kG_z) \\ &= (i \times i)F_x G_x + (i \times j)F_x G_y + (i \times k)F_x G_z \\ &\quad + (j \times i)F_y G_x + (j \times j)F_y G_y + (j \times k)F_y G_z \\ &\quad + (k \times i)F_z G_x + (k \times j)F_z G_y + (k \times k)F_z G_z. \end{aligned}$$

But now, as we readily see from the definition,

$$i \times i = j \times j = k \times k = 0,$$

(as, in fact, the vector product of any vector with itself is zero);
and

$$i \times j = -(j \times i) = k, \quad j \times k = -(k \times j) = i, \\ k \times i = -(i \times k) = j.$$

Hence, rearranging terms, we have

$$F \times G = i(F_y G_z - F_z G_y) + j(F_z G_x - F_x G_z) + \\ k(F_x G_y - F_y G_x). \quad (2)$$

As an example of the use of the vector product, we may mention the angular momentum vector. If we have, as in Chap. V, a particle of mass m , velocity v (a vector), and we wish its angular momentum about a certain center, we must take m times the magnitude of the radius vector times the projection of v at right angles to the radius. But this is just m times the magnitude of the vector product of r and v . Further, this vector product is a vector pointing along the axis of rotation, and in a positive direction if the rotation is positive, or counterclockwise, so that it is just in the direction conventionally assigned to the angular momentum. Hence we have angular momentum $= m(r \times v)$.

Another example of the use of the vector product comes frequently, when we may wish to prove two vectors to be parallel. To do this, we need only show that their vector product vanishes.

34. Vector Fields.—Very often in physics one has vectors which are functions of position. There are two particularly common examples, a force field, and a velocity, or flux density, in a flowing fluid. In an electric or magnetic or gravitational field, for instance, the force on unit charge or pole or mass at any point of space is a vector, of components F_x, F_y, F_z , varying from point to point in both direction and magnitude. Often such a vector field is indicated graphically by introducing lines tangent at every point to the vector at that point, called lines of force or lines of flow, as the case may be. We shall discuss the nature of vector fields more in detail in connection with hydrodynamics and the flow of fluids, in Chap. XVII. Our present application is to force fields, and our main interest is to discover in what cases the force vector can be derived from a potential

function. To investigate this, let us consider the energy theorem in three dimensions, deriving the work done in an arbitrary displacement.

35. The Energy Theorem in Three Dimensions.—Let us start with the equations of motion of a particle in a force field,

$$\begin{aligned} m \frac{d^2x}{dt^2} &= F_x, \\ m \frac{d^2y}{dt^2} &= F_y, \\ m \frac{d^2z}{dt^2} &= F_z. \end{aligned} \quad (3)$$

Multiplying by dx/dt , dy/dt , dz/dt , respectively, and integrating with respect to time, we have as in the last chapter

$$\begin{aligned} \frac{1}{2}m v_x^2 - \frac{1}{2}m v_{x0}^2 &= \int F_x dx, \\ \frac{1}{2}m v_y^2 - \frac{1}{2}m v_{y0}^2 &= \int F_y dy, \\ \frac{1}{2}m v_z^2 - \frac{1}{2}m v_{z0}^2 &= \int F_z dz. \end{aligned}$$

Adding,

$$\begin{aligned} \frac{1}{2}m (v_x^2 + v_y^2 + v_z^2) - \frac{1}{2}m (v_{x0}^2 + v_{y0}^2 + v_{z0}^2) \\ = \int (F_x dx + F_y dy + F_z dz). \end{aligned} \quad (4)$$

Now $v_x^2 + v_y^2 + v_z^2$ is the square of the magnitude of the velocity, or is v^2 . Thus the left side of Eq. (4) is the final kinetic energy of the particle minus the initial kinetic energy, so that the integral on the right should be the work done. The integrand is evidently a scalar product: the product of the vector F , and the infinitesimal displacement vector of components dx , dy , dz , which we may call ds . The scalar product, which is $F \cdot ds$, is the displacement times the projection of the force in the direction of motion. This is what is ordinarily called the work done, since only the component of force along the motion does work. The integral is simply the sum of all the infinitesimal amounts of work done, or is the total work done, as in one-dimensional motion.

36. Line Integrals and Potential Energy.—The integral $\int F \cdot ds$ is called a line integral, for its evaluation demands the knowledge of a definite path between starting point and end point, as well as of the function F . In general this integral will depend on the path as well as the end points. For instance, suppose the lines of force went in circles, as in Fig. 8. Then the work done along the path ABC is positive, since the force and

displacement are parallel; along ADC the work is negative, since force and displacement are opposite; while along AEC it is zero, force and displacement being at right angles. Intermediate paths would yield any value we chose for the work done. Hence we surely could not define a potential, for the work done between A and C could not be set equal to the difference of potential in any unique way. In discussing one-dimensional motion, we saw that a potential depending only on position could not be introduced if the force depended on velocity, time, or anything except displacement. Here the condition is more stringent: we cannot have a potential, even if force depends only on position, unless the integral $\int \mathbf{F} \cdot d\mathbf{s}$ is independent of path. If this condition is satisfied, however, we can set up a potential energy V , such that $-\int \mathbf{F} \cdot d\mathbf{s}$ from some standard point where the potential is zero, up to the point we are interested in, equals V . Evidently another way of stating the criterion for existence of

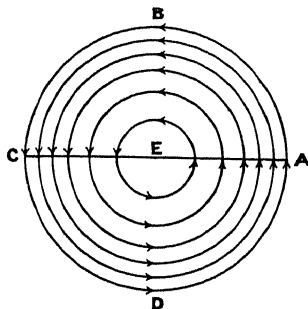


FIG. 8.—A nonconservative force field. The work done by the field on a particle moving along ABC is positive, along ADC negative, along AEC zero, so that the work done between A and C is not independent of path.

a potential is that the work done in taking a particle about any arbitrary closed path, or $\int \mathbf{F} \cdot d\mathbf{s}$ where the integral is about a closed curve and back to the starting point, be zero. Still a third condition, easier to apply in actual cases, will be derived in a later section.

37. Force as Gradient of Potential.—Let us suppose that it is possible to set up a potential function V in a given case. We know how to write V as the negative line integral of \mathbf{F} . Now we ask the opposite question: Given V , how do we find \mathbf{F} ? Let us suppose that we are at a given point of space, and that we allow the coordinates to increase by small amounts dx , dy , dz , forming a vector $d\mathbf{s}$, while at the same time we exert a force $-\mathbf{F}$ to balance the force of the field. Then first, we shall do the work $-\mathbf{F} \cdot d\mathbf{s}$ on the system; second, the potential will increase by the amount $dV = V(x + dx, y + dy, z + dz) - V(x, y, z)$. These must be equal; and writing the scalar product as $F_s |ds|$, where $|ds|$ is the magnitude of the displacement, F_s the component of \mathbf{F} parallel to the displacement, we have

$$dV = -F \cdot ds = -F_s |ds|, F_s = -\frac{dV}{|ds|}. \quad (5)$$

A derivative of the sort occurring in Eq. (5), where we take the difference of a scalar function like V at two neighboring points, divide by the magnitude of the displacement, and pass to the limit, is called a directional derivative, for evidently its value depends on the direction in which the displacement is made. We thus have the result that the component of force in any direction is the negative directional derivative of the potential in the desired direction.

The x component of force is determined from the directional derivative of V along the x direction. To find that, we allow x to increase by dx , keeping y and z fixed; divide the difference $V(x + dx, y, z) - V(x, y, z)$ by dx ; and pass to the limit as dx becomes small. But this is simply the partial derivative of V with respect to x . We see, in other words, that a partial derivative of a function is merely a special case of a directional derivative, in which the direction is along one of the coordinate axes. Using this fact, we then have

$$F_x = -\frac{\partial V}{\partial x}, F_y = -\frac{\partial V}{\partial y}, F_z = -\frac{\partial V}{\partial z}. \quad (6)$$

The three partial derivatives in Eq. (6) are evidently the components of a vector, called the gradient of V , and abbreviated $\text{grad } V$. Thus

$$\text{grad } V = i\frac{\partial V}{\partial x} + j\frac{\partial V}{\partial y} + k\frac{\partial V}{\partial z}, \quad (7)$$

and we may write a vector equation

$$F = -\text{grad } V. \quad (8)$$

38. Equipotential Surfaces.—Let us take a displacement ds in a direction tangent to an equipotential surface, or surface on which V is constant. Then no work is done, so that $dV = 0$. But also $F \cdot ds = 0$. If this is so, then F and ds must be at right angles. Thus we have proved that the force, and hence the lines of force, are at right angles to the equipotential surfaces. Any scalar function of position can be described by a set of surfaces, like equipotentials, on which it is constant. We see then that the gradient of such a function is a vector, at right angles to the

equipotentials, measuring the rate of change of the function in this direction. The name gradient comes from contour maps in two dimensions. There the contours are lines of constant altitude, and the ordinary gradient of a slope is the rate of change of height with horizontal distance, in the direction at right angles to the contours, or the direction of steepest slope. In our case, the gradient points in the direction in which the function increases, while the force, being the negative gradient of the potential, points in the direction in which the potential decreases.

39. The Curl and the Condition for a Conservative System.—Let $F_x = -\partial V/\partial x$, $F_y = -\partial V/\partial y$. Differentiating the first with respect to y , the second with respect to x , we have $\partial F_x/\partial y = -\partial^2 V/\partial y \partial x$, $\partial F_y/\partial x = -\partial^2 V/\partial x \partial y$. But by the fundamental theorem of partial differentiation, these two are equal, so that $\partial F_x/\partial y = \partial F_y/\partial x$. Similarly we have two other equations. These can be combined in a single vector equation. We shall find that it is useful to set up a vector called the curl, according to the definition

$$\text{curl } F = i\left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z}\right) + j\left(\frac{\partial F_x}{\partial z} - \frac{\partial F_z}{\partial x}\right) + k\left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y}\right). \quad (9)$$

Then our three equations are combined in the one vector equation $\text{curl } F = 0$. These form relations between the components of force, which plainly must be fulfilled if there is a potential. Yet it is by no means true that any set of forces will satisfy these conditions. The vanishing of the curl at all points of space, then, is a necessary condition which F must satisfy, if it is derivable from a potential. It can be proved that it is also a sufficient condition, so that it is the criterion which we desired, telling whether a potential can be set up in a given problem or not. As we shall see in a problem, the nonvanishing of the curl of a vector in general means whirlpool-like lines of force, as in Fig. 8.

40. The Symbolic Vector ∇ .—We have seen two vector differential operators, the gradient and the curl. These can both be expressed conveniently in terms of a symbolic vector operator ∇ , equal to $(i \partial/\partial x + j \partial/\partial y + k \partial/\partial z)$. Of course, this operator by itself has no meaning, but its interpretation is that it is always to be followed by some other quantity, and the differentiations are to be performed on this quantity. Thus if we have a scalar V , the quantity ∇V is a vector, equal to

$$\nabla V = \left(i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z} \right) V = i \frac{\partial V}{\partial x} + j \frac{\partial V}{\partial y} + k \frac{\partial V}{\partial z} = \text{grad } V. \quad (10)$$

Similarly, if we have a vector F , the vector product $(\nabla \times F)$ is equal to

$$(\nabla \times F) = i \left(\frac{\partial}{\partial y} F_z - \frac{\partial}{\partial z} F_y \right) + j \left(\frac{\partial}{\partial z} F_x - \frac{\partial}{\partial x} F_z \right) + k \left(\frac{\partial}{\partial x} F_y - \frac{\partial}{\partial y} F_x \right) = \text{curl } F. \quad (11)$$

In the course of time, we shall meet several other vector operations, which can be expressed in terms of ∇ . We shall merely define them now, though we shall have many applications later. If we have a vector F , the scalar product of ∇ with F , or $(\nabla \cdot F)$, is a scalar, evidently equal to

$$\nabla \cdot F = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z} = \text{div } F. \quad (12)$$

This is called the divergence of F , abbreviated $\text{div } F$. Again, if we have a scalar V , and take two factors ∇ multiplied by V , or $(\nabla \cdot \nabla)V$, the result is

$$\left(i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z} \right) \cdot \left(i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z} \right) V = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) V = \frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = \nabla^2 V. \quad (13)$$

This is called the Laplacian of V , and there is no usual abbreviation, except $\nabla^2 V$, which evidently is equivalent to the method of writing above. Clearly $\nabla^2 V = \text{div grad } V$. Finally we can take the Laplacian of a vector: if F is a vector,

$$\nabla^2 F = i \left(\frac{\partial^2 F_x}{\partial x^2} + \frac{\partial^2 F_x}{\partial y^2} + \frac{\partial^2 F_x}{\partial z^2} \right) + j \left(\frac{\partial^2 F_y}{\partial x^2} + \frac{\partial^2 F_y}{\partial y^2} + \frac{\partial^2 F_y}{\partial z^2} \right) + k \left(\frac{\partial^2 F_z}{\partial x^2} + \frac{\partial^2 F_z}{\partial y^2} + \frac{\partial^2 F_z}{\partial z^2} \right).$$

Problems

1. Find the angle between the diagonal of a cube and one of the edges. (Hint: regard the diagonal as a vector $i + j + k$.)

2. Given a vector $i + 2j + 3k$, and a second $i - 2j + ak$, find a so that the two vectors are at right angles to each other.

3. Let $F_x = y$, $F_y = -x$, $F_z = 0$. Prove that this vector field represents a force tangent to circles about the origin in the xy plane. Compute $\oint F \cdot ds$ around such a circle.

4. Find the curl of the force in the preceding problem. Discuss the question as to whether it is a conservative field or not.

5. In the gravitational field of a mass m , the potential is given by $-m/r$, where r is the distance from the mass, given by $r^2 = x^2 + y^2 + z^2$, if the mass is at the origin. Obtain the components of the force vector by direct differentiation. Find the curl of the force, and show that it is zero.

6. Find which ones of the following forces are derivable from potentials, and describe the physical nature of the force fields. Set up the potential in cases where that can be done:

$$(a) F_x = \frac{y}{x^2 + y^2}, F_y = \frac{-x}{x^2 + y^2}, F_z = 0.$$

$$(b) F_x = \frac{y}{\sqrt{x^2 + y^2}}, F_y = \frac{-x}{\sqrt{x^2 + y^2}}, F_z = 0.$$

(c) $F_x = xf(r)$, $F_y = yf(r)$, $F_z = zf(r)$, where $f(r)$ is an arbitrary function of the distance from the origin.

$$(d) F_x = f_1(x), F_y = f_2(y), F_z = f_3(z).$$

7. Prove that $lx + my + nz = k$, where l, m, n, k are constants, and $l^2 + m^2 + n^2 = 1$, is the equation of a plane whose normal has the direction cosines l, m, n , and whose shortest distance from the origin is k .

8. Taking the potential field from Prob. (5), find the line integral $\oint F \cdot ds$ around a square of arbitrary size in the xy plane, with the origin at its center. Show by direct calculation that the integral always vanishes. Do the same for a path made up as follows: the part of the square of side $2a$, made of lines at $x = -a$, $y = \pm a$, which lies at negative values of x , and the part of the circle of radius a , center at the origin, which joins onto and completes the figure for positive x 's.

9. Prove that $A \cdot (B \times C) = B \cdot (C \times A) = C \cdot (A \times B)$, where A, B, C are any vectors. Show that these are equal to the determinant

$$\begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix}.$$

10. Prove that $A \times (B \times C) = B(A \cdot C) - C(A \cdot B)$, where A, B, C are any vectors.

11. Prove that $\text{div } aF = a \text{ div } F + (F \cdot \text{grad } a)$, where a is a scalar, F a vector.

12. Prove that $\text{curl } aF = a \text{ curl } F + [(\text{grad } a) \times F]$, where a is a scalar, F a vector.

13. Prove that $\text{div } (F \times G) = (G \cdot \text{curl } F) - (F \cdot \text{curl } G)$, where F, G are vectors.

14. Prove that $\text{div curl } F = 0$, where F is any vector.

15. Prove that $\text{curl curl } F = \text{grad div } F - \nabla^2 F$, where F is any vector.

CHAPTER VII

LAGRANGE'S EQUATIONS AND PLANETARY MOTION

In considering mechanical problems with several variables, it is seldom very convenient to use ordinary rectangular coordinates. In working with problems in the motion of particles, we often wish to introduce curvilinear coordinates, as for instance polar coordinates. With rigid dynamics, we often use rather complicated quantities to give the orientation of a rigid body in space. For instance, with a top or gyroscope, we may use Euler's angles, namely, the latitude and longitude angles of the axis of the top with reference to a fixed north pole, and the angle of rotation of the top about its own axis. All these coordinates come under the general description of generalized coordinates. Any quantities which are capable of describing the positions of the parts of a system, whether they be distances, angles, or any other quantities, can serve as generalized coordinates. Now when we begin to examine the equations of motion in generalized coordinates, we naturally find that they can be very complicated. In a later section we shall start with the ordinary equations of motion in rectangular coordinates, introduce new coordinates as functions of the old, and find the new equations of motion by direct change of variables. We find many new terms coming in, as soon as the change of variables is at all complicated. But we shall find that there are several fairly simple ways of writing the equations of motion, different in form from Newton's equations, though essentially identical, which preserve their simple form even in generalized coordinates. The most elementary of these methods is that of Lagrange's equations, and we consider them in this chapter.

41. Lagrange's Equations.—We start our discussion of Lagrange's equations merely by restating Newton's second law of motion, in a slightly different way. For the moment, we consider only problems where there is a potential energy function. Since $F_x = -\partial V/\partial x$, etc., the equations of motion, written in terms of momenta, are

$$\frac{d(mv_x)}{dt} = -\frac{\partial V}{\partial x}, \quad (1)$$

etc. There is an interesting way in which these equations can be written. Let the kinetic energy be called T , so that

$$T = \frac{m}{2}(v_x^2 + v_y^2 + v_z^2), \quad (2)$$

if it is written in terms of the velocity components. If we keep this form, we observe that $mv_x = \partial T / \partial v_x$, which we note is the x component of momentum. Hence we can write our equations

$$\frac{d}{dt}\left(\frac{\partial T}{\partial v_x}\right) + \frac{\partial V}{\partial x} = 0, \quad (3)$$

etc. But this can be put in another form, if we let $T - V = L$, called the Lagrangian function (and different from the total energy, which is $T + V$). T is to be considered a function of the velocity components, and V of the coordinates, so that L is a function of all these six variables. Since T depends only on the v 's, and V only on the x 's, we have $\partial T / \partial v_x = \partial L / \partial v_x$, $\partial V / \partial x = -\partial L / \partial x$, etc. Hence the equations of motion are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial v_x}\right) - \frac{\partial L}{\partial x} = 0, \quad (4)$$

with similar equations for y and z . In this form, the equations are called Lagrange's equations of motion, and they are simply convenient ways of writing Newton's second law of motion.

As we have stated, the importance of Lagrange's equations is that they hold in any sort of coordinates, not merely in rectangular coordinates. Thus, if the coordinates are $q_1 \dots q_n$, and their time derivatives are $\dot{q}_1 \dots \dot{q}_n$, the equations are

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0. \quad (5)$$

Here as before $L = T - V$, but now it is no longer true, as before, that T depends only on the velocities, V only on the coordinates. Instead, T generally involves the coordinates as well, so that the term $\partial L / \partial q_i$ has some contributions coming from $\partial T / \partial q_i$, which are evidently absent in rectangular coordinates. We shall see by an example that these terms are a sort of fictitious force introduced by using the generalized coordinates,

and of which the centrifugal force in polar coordinates is a typical case. We postpone a proof of Lagrange's equations to a later section, giving first an example of their usefulness by discussing the motion of a particle in a central field, as a planet about the sun.

42. Planetary Motion.—As an example of two-dimensional motion, and of the Lagrangian equations, we consider the case where $V = V(r)$, a function only of the distance r from a given point. This problem is almost impossible to discuss completely if we use rectangular coordinates, but if we take polar coordinates, r, θ , we find that we can separate variables, and that the problem is then easily solved. To apply Lagrange's method to this case, we write L as a function of r, θ, \dot{r} , and $\dot{\theta}$. Then we have

$$\begin{aligned}\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}}\right) - \frac{\partial L}{\partial r} &= 0, \\ \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\theta}}\right) - \frac{\partial L}{\partial \theta} &= 0.\end{aligned}\tag{6}$$

First we find L . The velocity is made up of two vector components at right angles, along the radius and along the tangent to a circle. The first is \dot{r} , the second $r\dot{\theta}$, so that $v^2 = \dot{r}^2 + r^2\dot{\theta}^2$, and

$L = T - V = \frac{m}{2}(\dot{r}^2 + r^2\dot{\theta}^2) - V(r)$. Differentiating,

$$\begin{aligned}\frac{\partial L}{\partial \dot{r}} &= m\dot{r}, \\ \frac{\partial L}{\partial \dot{\theta}} &= mr^2\dot{\theta}, \\ \frac{\partial L}{\partial r} &= mr\dot{\theta}^2 - \frac{\partial V}{\partial r}, \\ \frac{\partial L}{\partial \theta} &= 0.\end{aligned}\tag{7}$$

Then the equations are

$$\begin{aligned}\frac{d}{dt}(m\dot{r}) - mr\dot{\theta}^2 + \frac{dV}{dr} &= 0, \\ \frac{d}{dt}(mr^2\dot{\theta}) &= 0.\end{aligned}\tag{8}$$

The second may be immediately integrated: $mr^2\dot{\theta} = \text{constant}$. This has a simple interpretation, for $mr^2\dot{\theta}$ is simply the angular momentum, since mr^2 is the moment of inertia, $\dot{\theta}$ the angular

velocity, and our equation states that it is constant, since no torque is acting. As a matter of fact, $\partial L / \partial \dot{q}_i$ is called the generalized momentum associated with the generalized coordinate q_i , and linear and angular momenta are special cases of the generalized momentum. Let then $mr^2\dot{\theta} = p$, where p is a constant (momenta are conventionally called p , as coordinates are called q). Next we may consider the first equation, $m \, d^2r/dt^2 = mr\dot{\theta}^2 - dV/dr$. The first term on the right-hand side is at first unexpected. But when we look at it, we see that it is the centrifugal force, which must be added to the external force to produce the radial acceleration.

We can now solve our equations. Setting $mr^2\dot{\theta} = p$, we have $\dot{\theta} = p/mr^2$, so that $m \, d^2r/dt^2 = p^2/mr^3 - dV/dr = -d/dr(V + p^2/2mr^2)$. We have separated the variable r from θ , and the result is just like the equation for a one-dimensional problem with a potential $V + p^2/2mr^2$, the latter being a sort of fictitious potential energy coming from the centrifugal force. For example, if the force is a gravitational one, $V = -Gmm'/r$, where m' is the mass of the attracting body, G the gravitational constant, so that we have the problem of the apparent potential $-Gmm'/r + p^2/2mr^2$. Except for the constants, this is the case of the potential $-(1/x) + (1/x^2)$, which we have already taken up in Probs. 3 and 4, Chap. V. We showed there that motions of negative energy are oscillatory in r , so that the orbit is concentrated in a finite region, and motions of positive energy go to infinity. We leave the exact discussion to a problem, but it proves to be true that the finite orbits are periodic and are ellipses with the attracting center at one focus, while the open orbits are hyperbolas. This is, however, a special case, and we proceed to a qualitative discussion of the general central motion, by the method of energy.

43. Energy Method for Radial Motion in Central Field.—We have seen that the radial motion of a particle in a central field is just like the one-dimensional motion of a particle in a potential $V + (p^2/2mr^2)$, where p is the constant angular momentum. This problem can be discussed as in Chap. V, plotting the curve $V + (p^2/2mr^2)$ as a function of r , and drawing the horizontal line at height E , as in Fig. 6. Aside from this, we can make no general statement. But in many important physical cases, the curve resembles *A* or *B* in Fig. 9, the rise at $r = 0$ arising from the centrifugal force, and the potential V representing attraction in *A*, repulsion in *B*. With energy E_1 , in either case, the motion

would come in from infinity to a smallest distance (c or d), called the perihelion, from the astronomical analogy, perihelion meaning near the sun. It would then reverse, and travel outward for infinite time. The energy E_2 , however, would represent no possible motion with the curve B , but with the attractive potential A , which resembles the gravitational attraction mentioned in the preceding section, there would be oscillatory motion between the perihelion a and the aphelion b . This motion

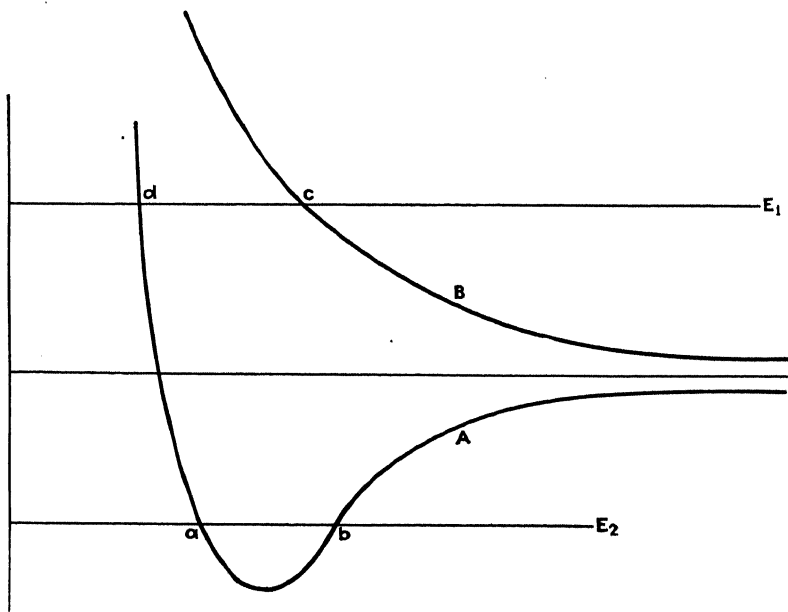


Fig. 9.—Curves of $V + \frac{p^2}{2mr^2}$ as functions of r . Case: A , attraction; B , repulsion. With energy E_1 , motion goes to infinity with either potential; with E_2 , motion impossible with curve B , oscillatory between limits a and b with curve A .

would be periodic, and the radius as function of time, and likewise the period, could be computed by the method of the energy integral discussed in Chap. V.

44. Orbits in Central Motion.—The best picture of central motion is obtained by considering the orbit in space, as in Fig. 10. Suppose we consider a motion oscillatory in r , as the case E_2 of Fig. 9. Then we may draw two circles, of radii equal to the perihelion and aphelion distances, respectively, and the motion will take place between the circles. The velocity must be

tangential to both circles, as shown. If the motion starts on the outer circle, the particle will move with continually decreasing radius until it touches the inner circle. At the same time, however, on account of the angular momentum, it will be turning around, and the angle made by the radius vector will have turned through a definite amount between the points of contact with outer and inner circles. After touching the inner circle, the whole procedure is reversed, r increasing to the maxi-

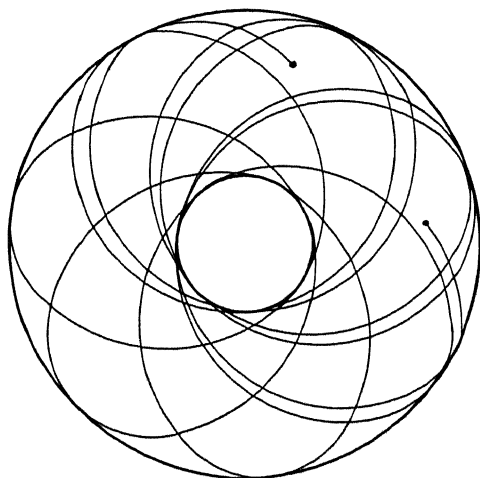


FIG. 10.—Orbit of a particle in central motion.

mum value, so that after a certain time the point will touch the outer circle again.

Now between the two successive points where the orbit touches the outer circle, there will be a certain length of arc. It may be that this is a rational fraction, say m/n , of the circumference, where m and n are integers. In that case, after n excursions to the center and out again, the aphelion point will have gone around the circle m times, and will have come back to the starting point. Thus the motion is periodic, repeating itself after a certain length of time. For example, if the particle is attracted to the center according to the inverse square, m/n is just 1, and the particle always comes back to the same point on the circle. But if the length of arc is an irrational fraction of the circumference, as in Fig. 10, the motion is not periodic, and will never repeat itself. Nevertheless, it is what is called doubly periodic. The motion resembles a slowly rotating ellipse,

rotating so that successive aphelion points, instead of lying on top of each other, are displaced with respect to each other by a given angle. This slow rotation is called precession, and one can find the frequency, and angular velocity, of the precessional motion. If now we imagined a turntable to rotate with the precessional frequency, and traced out the motion on this turntable, the path would be closed, somewhat like an ellipse. In other words, the whole motion is a combination of a periodic motion, superposed on a rotation. These two motions have in general entirely independent frequencies, and that is the origin of the statement that the motion is doubly periodic.

45. Justification of Lagrange's Method.—We shall now show in our special case of polar coordinates how Lagrange's method could be justified, using this as a model for the general treatment. Surely the equations of motion are

$$m \frac{d^2x}{dt^2} = -\frac{\partial V}{\partial x}, \quad m \frac{d^2y}{dt^2} = -\frac{\partial V}{\partial y}.$$

We introduce the polar coordinates, $x = r \cos \theta$, $y = r \sin \theta$. Then $dx/dt = \cos \theta \, dr/dt - r \sin \theta \, d\theta/dt$,

$$\frac{d^2x}{dt^2} = \frac{d^2r}{dt^2} \cos \theta - 2 \sin \theta \frac{dr}{dt} \frac{d\theta}{dt} - r \sin \theta \frac{d^2\theta}{dt^2} - r \cos \theta \left(\frac{d\theta}{dt} \right)^2, \quad (9)$$

$$\frac{d^2y}{dt^2} = \frac{d^2r}{dt^2} \sin \theta + 2 \cos \theta \frac{dr}{dt} \frac{d\theta}{dt} + r \cos \theta \frac{d^2\theta}{dt^2} - r \sin \theta \left(\frac{d\theta}{dt} \right)^2. \quad (10)$$

Using these, we can obtain the equations of motion in x and y . But now multiply Eq. (9) by $\bar{r} \cos \theta$, Eq. (10) by $\bar{r} \sin \theta$, and add. The result on the left is $m \, d^2r/dt^2 - mr(d\theta/dt)^2$, and on the right $-(\partial V/\partial x \cos \theta + \partial V/\partial y \sin \theta)$, which is just $-\partial V/\partial r$, since the latter should be $-(\partial V/\partial x \, \partial x/\partial r + \partial V/\partial y \, \partial y/\partial r)$, and $\partial x/\partial r = \cos \theta$, $\partial y/\partial r = \sin \theta$. Thus we have the first of Lagrange's equations. Next, multiply Eq. (9) by $-r \sin \theta$, Eq. (10) by $r \cos \theta$, and add. On the left, we have $2mr \, dr/dt \, d\theta/dt + mr^2 \, d^2\theta/dt^2$, which equals $m \, d/dt(r^2 d\theta/dt)$, and on the right we have $r \, \partial V/\partial x \sin \theta - r \, \partial V/\partial y \cos \theta = -\partial V/\partial \theta$. Thus the second equation becomes $m \, d/dt(r^2 d\theta/dt) = -\partial V/\partial \theta$, the second of Lagrange's equations (whose right member is zero in the case of a central field).

Just such a change of variables can be carried out in the general case. Suppose that, for the sake of simplicity, we still take only

two dimensions; the general proof goes through in just the same way, except with more complicated expressions. We start with two rectangular coordinates x and y , in terms of which we have the ordinary Newtonian equations $m \, d^2x/dt^2 = -\partial V/\partial x$, $m \, d^2y/dt^2 = -\partial V/\partial y$, and two generalized coordinates q_1 and q_2 , given as functions of x and y , so that $q_1 = q_1(x, y)$, $q_2 = q_2(x, y)$, or conversely we can write x and y as functions of q_1 and q_2 : $x = x(q_1, q_2)$, $y = y(q_1, q_2)$. We must remember carefully what these quantities are functions of, in taking partial derivatives. Now we have

$$\begin{aligned}\frac{dx}{dt} &= \frac{\partial x}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial x}{\partial q_2} \frac{dq_2}{dt}, \\ \frac{d^2x}{dt^2} &= \frac{\partial x}{\partial q_1} \frac{d^2q_1}{dt^2} + \frac{\partial x}{\partial q_2} \frac{d^2q_2}{dt^2} \\ &\quad + \frac{dq_1}{dt} \left(\frac{\partial^2 x}{\partial q_1^2} \frac{dq_1}{dt} + \frac{\partial^2 x}{\partial q_1 \partial q_2} \frac{dq_2}{dt} \right) \\ &\quad + \frac{dq_2}{dt} \left(\frac{\partial^2 x}{\partial q_1 \partial q_2} \frac{dq_1}{dt} + \frac{\partial^2 x}{\partial q_2^2} \frac{dq_2}{dt} \right)\end{aligned}$$

with a similar equation for d^2y/dt^2 . In terms of these, we set up the equations $m \, d^2x/dt^2 = -\partial V/\partial x$, etc. Then we multiply the first by $\partial x/\partial q_1$, the second by $\partial y/\partial q_1$, and add. We have

$$\begin{aligned}m \left\{ \left[\left(\frac{\partial x}{\partial q_1} \right)^2 + \left(\frac{\partial y}{\partial q_1} \right)^2 \right] \frac{d^2q_1}{dt^2} + \left(\frac{\partial x}{\partial q_1} \frac{\partial x}{\partial q_2} + \frac{\partial y}{\partial q_1} \frac{\partial y}{\partial q_2} \right) \frac{d^2q_2}{dt^2} \right. \\ \left. + \left(\frac{\partial x}{\partial q_1} \frac{\partial^2 x}{\partial q_1^2} + \frac{\partial y}{\partial q_1} \frac{\partial^2 y}{\partial q_1^2} \right) \left(\frac{dq_1}{dt} \right)^2 \right. \\ \left. + 2 \left(\frac{\partial x}{\partial q_1} \frac{\partial^2 x}{\partial q_1 \partial q_2} + \frac{\partial y}{\partial q_1} \frac{\partial^2 y}{\partial q_1 \partial q_2} \right) \frac{dq_1}{dt} \frac{dq_2}{dt} \right. \\ \left. + \left(\frac{\partial x}{\partial q_1} \frac{\partial^2 x}{\partial q_2^2} + \frac{\partial y}{\partial q_1} \frac{\partial^2 y}{\partial q_2^2} \right) \left(\frac{dq_2}{dt} \right)^2 \right\} \\ = - \left(\frac{\partial V}{\partial x} \frac{dx}{\partial q_1} + \frac{\partial V}{\partial y} \frac{dy}{\partial q_1} \right) = - \frac{\partial V}{\partial q_1}.\end{aligned}$$

It will next be shown that the rather complicated expression on the left is equal to

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_1} \right) - \frac{\partial T}{\partial q_1},$$

where T is the kinetic energy. To do this, we first have

$$T = \frac{m}{2} \left[\left(\frac{\partial x}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial x}{\partial q_2} \frac{dq_2}{dt} \right)^2 + \left(\frac{\partial y}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial y}{\partial q_2} \frac{dq_2}{dt} \right)^2 \right].$$

Then by differentiation, remembering that $\dot{q}_1 = dq_1/dt$,

$$\begin{aligned} \frac{\partial T}{\partial \dot{q}_1} &= m \left[\left(\frac{\partial x}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial x}{\partial q_2} \frac{dq_2}{dt} \right) \frac{\partial x}{\partial q_1} + \left(\frac{\partial y}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial y}{\partial q_2} \frac{dq_2}{dt} \right) \frac{\partial y}{\partial q_1} \right] \\ \frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_1} \right) &= m \left[\left(\frac{\partial x}{\partial q_1} \frac{d^2 q_1}{dt^2} + \frac{\partial x}{\partial q_2} \frac{d^2 q_2}{dt^2} \right) \frac{\partial x}{\partial q_1} + \left(\frac{\partial y}{\partial q_1} \frac{d^2 q_1}{dt^2} + \frac{\partial y}{\partial q_2} \frac{d^2 q_2}{dt^2} \right) \frac{\partial y}{\partial q_1} \right] \\ &+ m \left\{ \left(\frac{dq_1}{dt} \right)^2 \left[\frac{\partial}{\partial q_1} \left(\frac{\partial x}{\partial q_1} \right)^2 + \frac{\partial}{\partial q_1} \left(\frac{\partial y}{\partial q_1} \right)^2 \right] \right. \\ &+ \frac{dq_1}{dt} \frac{dq_2}{dt} \left[\frac{\partial}{\partial q_2} \left(\frac{\partial x}{\partial q_1} \right)^2 + \frac{\partial}{\partial q_2} \left(\frac{\partial y}{\partial q_1} \right)^2 + \frac{\partial}{\partial q_1} \left(\frac{\partial x}{\partial q_1} \frac{\partial x}{\partial q_2} \right) + \frac{\partial}{\partial q_1} \left(\frac{\partial y}{\partial q_1} \frac{\partial y}{\partial q_2} \right) \right] \\ &\left. + \left(\frac{dq_2}{dt} \right)^2 \left[\frac{\partial}{\partial q_2} \left(\frac{\partial x}{\partial q_1} \frac{\partial x}{\partial q_2} \right) + \frac{\partial}{\partial q_2} \left(\frac{\partial y}{\partial q_1} \frac{\partial y}{\partial q_2} \right) \right] \right\}. \end{aligned}$$

Also

$$\begin{aligned} \frac{\partial T}{\partial q_1} &= m \left[\left(\frac{\partial x}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial x}{\partial q_2} \frac{dq_2}{dt} \right) \left(\frac{\partial^2 x}{\partial q_1^2} \frac{dq_1}{dt} + \frac{\partial^2 x}{\partial q_1 \partial q_2} \frac{dq_2}{dt} \right) \right. \\ &\left. + \left(\frac{\partial y}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial y}{\partial q_2} \frac{dq_2}{dt} \right) \left(\frac{\partial^2 y}{\partial q_1^2} \frac{dq_1}{dt} + \frac{\partial^2 y}{\partial q_1 \partial q_2} \frac{dq_2}{dt} \right) \right] \end{aligned}$$

Combining these two expressions, it is easy to see that we have just the quantity which we desired. We have then the equation

$$\frac{d}{dt} \left(\frac{\partial T}{\partial \dot{q}_1} \right) - \frac{\partial T}{\partial q_1} = - \frac{\partial V}{\partial q_1}.$$

If we set $L = T - V$, and remember that, since V does not depend on the velocities, $\partial V / \partial \dot{q}_1 = 0$, this becomes

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_1} \right) - \frac{\partial L}{\partial q_1} = 0,$$

or Lagrange's equation for q_1 . Similarly we can prove the equation for q_2 .

It is worth remarking that the method which we have used for proving Lagrange's equations, though straightforward and simple in principle, is not the one usually employed. More often a derivation is given using the calculus of variations, which avoids most of the algebraic complications, but which on the other hand is more difficult in the fundamental ideas involved.

Problems

1. A particle of mass m is attracted to a center by a force $-Gmm'/r^2$. Find perihelion and aphelion distances as a function of energy and angular momentum. Assuming that the orbit is an ellipse, prove that its major axis is $-Gmm'/E$.

2. In Prob. 1, show that it is possible for perihelion and aphelion distances to be equal, so that the orbit is circular. Find the necessary relation between energy and angular momentum for this to happen, and check this relation by elementary discussion, balancing the centrifugal force in the circular motion against the attraction.

3. A particle in an inverse square field executes an elliptical motion with the center of attraction as a focus. Find the period of this motion, by considering the radial motion, proceeding as in Prob. 5, Chap. V, using the results of that problem if you wish, but finding the period in terms of energy and angular momentum.

4. Discuss in detail the motion of a planet about a sun, proving that, if the energy is negative, the orbit is elliptical with the sun at a focus, and finding the relations between the major and minor axes of the ellipse and the energy and angular momentum. A procedure for the discussion is suggested as follows:

Assuming the angular momentum to be $p = mr^2\dot{\theta} = \text{constant}$, show that the energy is $\frac{p^2}{2m} \left[\left(\frac{du}{d\theta} \right)^2 + u^2 \right] - Gmm'u$, where $u = \frac{1}{r}$. Find $\frac{du}{d\theta}$ from the equation of an ellipse in polar coordinates, with one focus as a pole, which is $u = \frac{1 - \epsilon \cos \theta}{a(1 - \epsilon^2)}$, where a is the semi-major axis, ϵ the eccentricity, so that b , the semi-minor axis, is given by $b^2/a^2 = 1 - \epsilon^2$. Substituting your value of $du/d\theta$ into the expression for energy, show that the result is a constant, independent of θ , and equal to E , if the major axis and eccentricity are properly chosen.

5. Suppose a particle of mass m , charge e , collides with a very heavy particle which has charge e' , so that it repels with a potential energy ee'/r . The first particle is moving with a velocity v_0 at a great distance, and is aimed so that, if it continued in a straight line, it would pass by the center of repulsion at a minimum distance R . Note that this determines the angular momentum. Using the energy method, find the perihelion distance as a function of R and the velocity of the particle.

6. Discuss in detail the motion of the particle of Prob. 5, showing that it will be deflected so that after the collision the line of travel will make an angle ϕ with the initial direction, where $\tan \frac{\phi}{2} = \frac{ee'}{mv_0^2 R}$. Such deflections are observed in collisions between alpha particles and atomic nuclei, in Rutherford's scattering experiments.

Suggestions: the particle executes a hyperbolic orbit, and the desired angle is the angle between the asymptotes. Now the equation of a hyperbola in polar coordinates is just like that of an ellipse, as given in Prob. 4, except that the eccentricity is greater than 1, so that the term $1 - \epsilon \cos \theta$ can become zero, and r infinite, giving the angles of the asymptotes in

terms of ϵ . We need then only determine ϵ in terms of energy and angular momentum, from the equations found in Prob. 4.

7. A two-dimensional linear oscillator is attracted to a center by a force proportional to the distance, or $F_x = -ax$, $F_y = -ay$. Solve in rectangular coordinates, separating variables, showing that x and y execute independent simple harmonic vibrations of the same frequency. Prove that the resulting orbit is an ellipse, with its center at the center of attraction.

8. Taking the solution of Prob. 7 in rectangular coordinates, find the angular momentum vector by ordinary vector formulas from the displacement and velocity, and prove by direct computation that it remains constant. Find the angular momentum as a function of the dimensions of the elliptical orbit, and show its connection with the area of the orbit.

9. Set up the problem of the two-dimensional linear oscillator, as in Prob. 7, using polar coordinates. Separate variables, solve the radial problem by the energy method, compute the period in this way, and show that it is in agreement with the period as found in Prob. 7.

CHAPTER VIII

GENERALIZED MOMENTA AND HAMILTON'S EQUATIONS

In the last chapter we have found the equations of motion in generalized coordinates, but we have not considered the meaning in these coordinates of the simple concepts of momentum and force. We shall accordingly examine these questions, and shall see that the equations can be interpreted in the form that the force equals the time rate of change of momentum, which as we know is a more fundamental statement than that it is the mass times acceleration. Using the momentum, we can then restate the equations in a form called Hamilton's equations, equivalent to Lagrange's equations, but more powerful in some applications to advanced mechanics.

46. Generalized Forces.—In many mechanical problems we have to deal with forces which cannot be derived from a potential. Let us see how such forces may be included in the Lagrangian scheme. For simplicity we take a two-dimensional problem, and let the x and y components of force be F_x and F_y , which may depend on time, velocity, etc., as well as position. For generality, we assume that part of the force can be derived from a potential, the rest not, so that we have $F_x = -(\partial V/\partial x) + F_x'$, etc., where F_x' is the part of the force not derivable from a potential. Now if we proceed with the proof of Lagrange's equations as in the last chapter, we easily find

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_1}\right) - \frac{\partial T}{\partial q_1} = -\frac{\partial V}{\partial q_1} + \left(F_x' \frac{\partial x}{\partial q_1} + F_y' \frac{\partial y}{\partial q_1}\right),$$

with a similar equation for q_2 . We may introduce as before a Lagrangian function, containing the part of the external forces derivable from a potential: $L = T - V$. Then

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_1}\right) - \frac{\partial L}{\partial q_1} = F_x' \frac{\partial x}{\partial q_1} + F_y' \frac{\partial y}{\partial q_1} = Q_1, \quad (1)$$

with a similar equation for q_2 , where Q_1, Q_2 are called the gen-

eralized forces connected with the coordinates q_1, q_2 . The equation in this form may be used to discuss any arbitrary problem, for example of damped motion, in generalized coordinates.

It is worth noting that these generalized forces are closely related to the work done in an arbitrary displacement, just as ordinary forces are in rectangular coordinates. For imagine the generalized coordinates changed by amounts dq_1, dq_2 . There will be a certain amount of work done on the system, equal to $-dV + dW$, where dW is the work done by the external non-conservative force F' (a force is spoken of as conservative if it is derivable from a potential, nonconservative otherwise). Now in general we have

$$\begin{aligned} dW &= F'_x dx + F'_y dy \\ &= F'_x \left(\frac{\partial x}{\partial q_1} dq_1 + \frac{\partial x}{\partial q_2} dq_2 \right) + F'_y \left(\frac{\partial y}{\partial q_1} dq_1 + \frac{\partial y}{\partial q_2} dq_2 \right) \\ &= \left(F'_x \frac{\partial x}{\partial q_1} + F'_y \frac{\partial y}{\partial q_1} \right) dq_1 + \left(F'_x \frac{\partial x}{\partial q_2} + F'_y \frac{\partial y}{\partial q_2} \right) dq_2 \\ &= Q_1 dq_1 + Q_2 dq_2, \end{aligned} \quad (2)$$

or the sum of products of generalized forces by generalized displacements. It is, of course, plain that all these arguments work equally well with more than two generalized coordinates.

The forces Q which we have just introduced were the external applied forces not derivable from a potential. But we may well consider all the forces together. We could write Lagrange's equations as

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = Q_i - \frac{\partial V}{\partial q_i} + \frac{\partial T}{\partial q_i}. \quad (3)$$

The three terms on the right of Eq. (3) may be taken to be three terms of the force. The first is the generalized force not derivable from a potential, the second the force derivable from a potential, the third the fictitious force, like a centrifugal force, arising from the fact that the coordinate system is not rectangular. Equation (3) states that this total force equals the time rate of change of a certain quantity, and it seems reasonable to consider this quantity as a generalized momentum.

47. Generalized Momenta.—In simple cases the quantity $\partial L / \partial \dot{q}_i$ plays the part of a momentum. Thus in rectangular coordinates, we have $\partial L / \partial \dot{x} = m\dot{x}$, or exactly the momentum associated with the coordinate x . Similarly in polar coordinates

the quantities associated with r and θ are $m\dot{r}$, the radial momentum, and $m\dot{r}^2\dot{\theta}$, the angular momentum, respectively. These are but examples of a general rule, and as a matter of fact we define $\partial L/\partial \dot{q}_i$ to be the generalized momentum associated with the coordinate q_i , denoting it by p_i . We note that generalized momenta are not of the same dimensions as ordinary momenta, in general; they are not simply components of the momentum referred to other coordinates. Similarly generalized forces are not simply components of forces. For instance, it is easily shown that in polar coordinates the generalized force Q_r is the component of force along r , but Q_θ is the moment of force, or torque, which by Eq. (3) above equals the time rate of change of the angular momentum.

48. Hamilton's Equations of Motion.—Assuming no external forces Q , we could evidently write Lagrange's equations in the form $dp_i/dt - \partial L/\partial q_i = 0$, or $dp_i/dt = \partial L/\partial q_i$, which, taken together with the definitions $p_i = \partial L/\partial \dot{q}_i$, would form a complete system. But there is a neater method, known as Hamilton's method, which we use instead. We can first see how Hamilton's equations are set up in rectangular coordinates. There we have $T = (m/2)(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$. Then it is true that we have, for instance,

$$p_x = \frac{\partial L}{\partial \dot{x}} = \frac{\partial T}{\partial \dot{x}} = m\dot{x}.$$

We can also write T , not in terms of the velocities \dot{x} , \dot{y} , \dot{z} , but in terms of the momenta p_x , p_y , p_z . Since $\dot{x} = p_x/m$, we have

$$T(p_x, p_y, p_z) = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2),$$

where we must specify that T is a function of the p 's. Then we have

$$\frac{\partial T(p_x, p_y, p_z)}{\partial p_x} = \frac{p_x}{m} = \frac{m\dot{x}}{m} = \dot{x},$$

and similarly $\partial T(p)/\partial p_y = \dot{y}$, $\partial T(p)/\partial p_z = \dot{z}$. These take the place of the equations $p_x = \partial T(\dot{x}, \dot{y}, \dot{z})/\partial \dot{x}$, etc.

Now in Hamilton's method we set up what is called the Hamiltonian function H . This is in all ordinary cases simply the total energy $T + V$, in which T is expressed in terms of the momenta, rather than the velocities. Thus we have $H = H(q_i, p_i)$, mean-

ing that it is a function of the coordinates and momenta. Then

$$\frac{\partial H}{\partial q_i} = \frac{\partial T}{\partial q_i} + \frac{\partial V}{\partial q_i},$$

which in rectangular coordinates gives $\partial V/\partial q_i = -\partial L/\partial q_i$, so that in this case Lagrange's equation becomes $dp_i/dt = -\partial H/\partial q_i$. Similarly

$$\frac{\partial H}{\partial p_i} = \frac{\partial T}{\partial p_i} = \dot{q}_i = \frac{dq_i}{dt}.$$

The resulting equations are called Hamilton's equations:

$$\begin{aligned}\frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}\end{aligned}\tag{4}$$

It is evident that they show a symmetry between p_i and q_i , which is one reason for preferring them over Lagrange's equations. For a given problem, there are twice as many Hamiltonian equations as Lagrangian equations, but they are only first-order rather than second-order differential equations, so that it comes down essentially to the same thing.

49. General Proof of Hamilton's Equations.—Our proof holds only in rectangular coordinates, and we must next give a general proof. As before, we start with Lagrange's equations, which we assume are correct, and we define the momenta as derivatives of the Lagrangian function with respect to the velocities. Then we set up the Hamiltonian function in terms of the Lagrangian function, by the equation

$$H = \sum_j p_j \dot{q}_j - L.\tag{5}$$

This seems at first quite different from our elementary definition of H as the energy, but we shall show in the next paragraph that it is equivalent. We express the Hamiltonian in terms of coordinates and momenta, writing the velocities \dot{q}_i , where they appear both in $\sum p_j \dot{q}_j$ and L , in terms of the momenta, so that we have

$$H = \sum_j p_j \dot{q}_j(p_k, q_k) - L[\dot{q}_j(p_k, q_k), q_j].$$

Then we have

$$\frac{\partial H}{\partial p_i} = \dot{q}_i + \sum_j p_j \frac{\partial \dot{q}_j}{\partial p_i} - \sum_j \frac{\partial L}{\partial \dot{q}_j} \frac{\partial \dot{q}_j}{\partial p_i}.$$

But since by definition $p_j = \partial L / \partial \dot{q}_j$, the last two terms cancel, leaving

$$\frac{\partial H}{\partial p_i} = \dot{q}_i = \frac{dq_i}{dt}.$$

Similarly,

$$\frac{\partial H}{\partial q_i} = \sum_j p_j \frac{\partial \dot{q}_j}{\partial q_i} - \sum_j \frac{\partial L}{\partial \dot{q}_j} \frac{\partial \dot{q}_j}{\partial q_i} - \frac{\partial L}{\partial q_i}.$$

This time the first two terms cancel, leaving $\partial H / \partial q_i = -\partial L / \partial q_i$, so that by Lagrange's equations.

$$\frac{\partial H}{\partial q_i} = -\frac{dp_i}{dt}.$$

Thus we have proved both of Hamilton's equations in the general case.

It remains to be shown that the Hamiltonian function, as we have defined it, is the same as the total energy. First we consider the kinetic energy expressed in terms of the velocities. This is a homogeneous quadratic function of the velocities:

$$T = \sum_{jk} A_{jk} \dot{q}_j \dot{q}_k, \quad (6)$$

where the A 's are coefficients depending in general on the coordinates, and we are to sum over all possible values j and k . In particular, for rectangular coordinates, $A_{jk} = m/2$ if $j = k$, 0 if $j \neq k$. In cases where the coordinates are orthogonal, that is, the coordinate surfaces intersect at right angles, as they do, for instance, in spherical polar coordinates, or in fact in all the coordinate systems in common use, only square terms come in, all coefficients A_{jk} being zero if $j \neq k$. But in oblique coordinate systems, this is not true. Now for such a homogeneous quadratic expression we have the theorem

$$2T = \sum_i \frac{\partial T}{\partial \dot{q}_i} \dot{q}_i,$$

which we can immediately prove. For $\frac{\partial T}{\partial \dot{q}_i} = \sum_j (A_{ji} + A_{ij}) \dot{q}_j$,

so that

$$\sum_i \frac{\partial T}{\partial \dot{q}_i} \dot{q}_i = \sum_i \sum_j (A_{ij} + A_{ji}) \dot{q}_i \dot{q}_j.$$

The double sum is now just twice the sum of Eq. (6) which we previously gave as T , proving the theorem. Hence, using $\partial T / \partial \dot{q}_i = \partial L / \partial \dot{q}_i = p_i$, we have $T = \frac{1}{2} \sum_i p_i \dot{q}_i$, so that our defini-

tion in Eq. (5) of H gives $H = 2T - L = 2T - T + V = T + V = \text{total energy}$, as we wished to prove.

In advanced work, one sometimes meets cases where H is not equivalent to the total energy. Such cases are found, for instance, where magnetic forces are present. But even here, the following general rules are correct:

First, set up a Lagrangian function, so that the equations of motion can be written in Lagrangian form. This can sometimes, as in the magnetic case, be done, even if we cannot interpret the Lagrangian function as $T - V$; for in the magnetic case, the forces are not derivable from a potential, depending rather on the velocity, and yet vary in such a way that we can use a Lagrangian function.

Next, define the momenta as $p_i = \partial L / \partial \dot{q}_i$.

Set up the Hamiltonian function $\sum p_i \dot{q}_i - L$, expressing it in terms of coordinates and momenta.

Then Hamilton's equations hold, using this Hamiltonian.

50. Example of Hamilton's Equations.—Let us by way of illustration work out Hamilton's equations for the problem of planetary motion, discussed in the previous chapter by Lagrange's method. In terms of the coordinates r and θ , we found that

$$L = \frac{m}{2}(\dot{r}^2 + r^2\dot{\theta}^2) - V(r).$$

Then the momenta are $p_r = \partial L / \partial \dot{r} = m\dot{r}$, the ordinary momentum along the radius, and $p_\theta = \partial L / \partial \dot{\theta} = mr^2\dot{\theta}$, the angular momentum. Next we have

$$\begin{aligned} \sum p_i \dot{q}_i - L &= (m\dot{r})\dot{r} + (mr^2\dot{\theta})\dot{\theta} - L \\ &= m(\dot{r}^2 + r^2\dot{\theta}^2) - \frac{m}{2}(\dot{r}^2 + r^2\dot{\theta}^2) + V(r) \end{aligned}$$

$$= \frac{m}{2}(\dot{r}^2 + r^2\dot{\theta}^2) + V(r)$$

= total energy.

Solving the equations for \dot{r} and $\dot{\theta}$ in terms of p_r and p_θ , we have $\dot{r} = p_r/m$, $\dot{\theta} = p_\theta/mr^2$, and substituting these in the Hamiltonian, we have

$$H = \frac{1}{2m}\left(p_r^2 + \frac{1}{r^2}p_\theta^2\right) + V(r).$$

Then Hamilton's equations are

$$\begin{aligned}\frac{\partial H}{\partial p_r} &= \frac{p_r}{m} = \frac{dr}{dt} = \dot{r} \\ \frac{\partial H}{\partial p_\theta} &= \frac{p_\theta}{mr^2} = \frac{d\theta}{dt} = \dot{\theta},\end{aligned}$$

both of which we already knew. Also

$$-\frac{\partial H}{\partial r} = \frac{p_\theta^2}{mr^3} - \frac{\partial V(r)}{\partial r} = \frac{dp_r}{dt},$$

showing that the time rate of change of radial momentum equals the external force $-\partial V/\partial r$ in the r direction, plus the centrifugal force p_θ^2/mr^3 (which evidently equals $mr\dot{\theta}^2 = m\omega^2 r = mv^2/r$). Finally

$$-\frac{\partial H}{\partial \theta} = 0 = \frac{dp_\theta}{dt},$$

showing that the time rate of change of angular momentum is zero, on account of the absence of torques.

51. Applications of Lagrange's and Hamilton's Equations.—From our discussion one might get the impression that the only use of Lagrange's and Hamilton's equations was in introducing curvilinear coordinates in problems of the dynamics of a particle. This is, however, far from the case. For example, one may have a particle moving subject to certain constraints, as a bead sliding along a frictionless wire, or a particle constrained to move on the surface of a sphere or other surface, as the bob of a spherical pendulum must move in a sphere. Then we may often satisfy the conditions of constraint by suitable choice of the generalized coordinates. Thus, with the spherical pendulum, we may take spherical polar coordinates r , θ , ϕ . We may then arbitrarily set r constant, equal to R , the radius of the sphere, and write

Lagrange's equations for θ and ϕ . To justify this, we note that the component of the external and centrifugal force normal to the sphere will be exactly balanced by the reaction of the constraint, just as the weight of a body resting on a table is exactly balanced by the upward push of the table. Thus the generalized force acting in the direction of r will be zero, so that a constant value for R , leading to a constant and vanishing generalized momentum along r , is a solution of the equations. For a particle on a wire, similarly, if the wire happened to be a circle, we could take polar coordinates, set r constant, and have but one equation of motion, stating that the torque acting on the particle equaled the time rate of change of its angular momentum. We note that these two problems are essentially equivalent to the spherical and ordinary pendulum, which are rigid bodies, suggesting that Lagrange's equations are of use in discussing the motion of a rigid body. But we can go even further. An Atwood's machine, for instance, is a special case of coupled systems, two weights being hung by a string over a pulley. This can be described very easily by a single generalized coordinate. In the general problem of coupled systems, and in fact in all problems of interaction of different particles or systems, Lagrange's method is very suitable, as we shall see. In fact, there is hardly a mechanical problem where generalized coordinates are not applied.

For the actual solution of problems, Hamilton's equations are generally not so convenient as Lagrange's equations. Their importance comes in the insight they give into the situation, by bringing the momenta directly into the statement of the equations, and for their relation to more advanced mechanics. The applications are principally to three fields: celestial mechanics, statistical mechanics, and quantum theory. We shall indicate in the next chapter the nature of some of these applications of Hamiltonian methods, taking up some of the general properties of the motion of particles, but postponing until later in the book the discussion of statistics and of quantum mechanics.

Problems

1. An Atwood's machine is built as follows: A string of length l_1 passes over a light fixed pulley, supporting a mass m_1 on one end and a pulley of mass m_2 (negligible moment of inertia) on the other. Over this second pulley passes a string of length l_2 supporting a mass m_3 on one end and m_4 on the other, where $m_3 \neq m_4$. Set up Lagrange's equations of motion for this

system, using two appropriate generalized coordinates. From these show that the mass m_1 remains in equilibrium if

$$m_1 = m_2 + m_3 + m_4 - \frac{(m_4 - m_3)^2}{m_3 + m_4}.$$

2. A particle slides on the inside of a smooth paraboloid of revolution whose axis is vertical. Use the distance from the axis, r , and the azimuth θ as generalized coordinates. Find the equations of motion. Find the angular momentum necessary for the particle to move in a horizontal circle. If this latter motion is disturbed slightly, show that the particle will perform small oscillations about this circular path, and find the period of these oscillations.

3. Set up the kinetic energy, Lagrange's equations, and Hamilton's equations in spherical polar coordinates. Set up expressions for the generalized forces acting on r , θ , and ϕ , and for the generalized momenta, explaining the physical meaning of these quantities.

4. Set up the problem of a spherical pendulum subject to gravity and to a resisting force proportional to the velocity and opposite in direction. Use spherical polar coordinates. Show that for small amplitudes and no damping this problem reduces to the two-dimensional linear oscillator of Prob. 9, Chap. VII.

5. Derive the Hamiltonian equations for Prob. 4, in the general case, showing that the damping forces give extra terms in the equations proportional to the momenta. Show that these equations in general cannot be separated. Derive a solution, however, for the special case in which the instantaneous motion would be a rotation about the lowest point of the sphere if damping were absent. Assume small damping, so that the actual motion is a gradual spiralling in toward the lowest point.

6. The force on an electron of charge e , moving with a velocity v in a magnetic field H , is given by $F = \frac{e}{c}(v \times H)$, where c is the velocity of light.

This corresponds to the ordinary motor law, in which the force on a circuit is proportional to the current (here ev/c) and to the field, and at right angles to both. In addition, the magnetic field H can be given as the curl of a vector A , called the vector potential. Show that the equations of motion of an electron moving in such a magnetic field, and in addition in a potential field of potential V , can be described by Lagrange's equations, with the Lagrangian function $L = T - V + (e/c)(v \cdot A)$. Assume the vector potential, and magnetic field, to be independent of time, but note that

$$\frac{dA}{dt} = \frac{\partial A}{\partial t} + \frac{\partial A}{\partial x} \frac{dx}{dt} + \frac{\partial A}{\partial y} \frac{dy}{dt} + \frac{\partial A}{\partial z} \frac{dz}{dt}, \text{ where } \frac{\partial A}{\partial t} = 0.$$

7. For the particle of Prob. 6, set up the momentum and the Hamiltonian function. Show that the momenta do not equal mass times velocity, and the Hamiltonian is not the same as the total energy.

8. In the relativity theory, the equations of motion of a particle are different from what they are in classical mechanics, though they reduce to the same thing for small velocities. In particular, the mass of a particle

increases with velocity. If a particle has a mass m_0 when at rest, its mass at speed v is given by

$$m = \frac{m_0}{\sqrt{1 - v^2/c^2}},$$

where c is the velocity of light; reducing to m_0 in the limit $v/c = 0$, but becoming infinite when the particle moves with the speed of light.

Show that the equations of motion are correctly given from the Lagrangian function $-m_0c^2\sqrt{1 - v^2/c^2} - V$, when we remember that the momentum equals the velocity times the (variable) mass.

Derive the Hamiltonian function from the Lagrangian function. Setting the Hamiltonian function equal to $T + V$, where T is the kinetic energy, show that the Lagrangian function is not equal to $T - V$, as is natural from the fact that the kinetic energy is not a homogeneous quadratic function of the velocities. Taking the kinetic energy, expand in power series in the quantity v/c , showing that for low speeds the kinetic energy approaches its ordinary classical value, except for an additive constant m_0c^2 . This additive constant, which always appears in relativistic energy expressions, is interpreted as meaning that the mass of the particle is really equivalent to energy, 1 gm. being convertible into c^2 ergs of energy.

CHAPTER IX

PHASE SPACE AND THE GENERAL MOTION OF PARTICLES

As in one-dimensional motion, we can make a great deal of use of the energy in discussing motion in two and three dimensions. In a conservative system with potential energy V , the motion can occur only in those regions of space where $E - V$ is positive, if E is the total energy, and we can thus divide up our possible problems into those occurring within a finite region and those going to infinity. As with one-dimensional motion, there are sometimes periodicity properties associated with the finite motions, which we discuss in the present chapter. With two-dimensional motion, we can visualize the use of the energy very easily, plotting V as a height in a three-dimensional graph, the result looking like a relief map, or else drawing equipotentials, which represent the potential as the contour lines represent height on a map. For a total energy E , we imagine the map filled with water up to a level E , so that the submerged parts, lakes and oceans, represent the regions where the motion occurs. We may also use the analogy of the rolling ball in two dimensions as well as in one, imagining that a ball starts rolling down the side of a hill in our relief map, climbing up the valley on the other side, and oscillating back and forth. From physical intuition as to the motion of such a ball, we can derive much information about complicated forms of motion.

There is one great complication present in motion in several dimensions which was absent in one-dimensional motion. In that simpler case, the velocity of a particle was determined at each point of space in a conservative motion, only the direction, forward or back, being arbitrary. Here, however, while the magnitude of the velocity is still determined, there are an infinite number of possible directions associated with the same magnitude. To describe a motion completely, then, even if we know its energy, we must give as well the velocity components, or else the momenta, at each point of the path. This is accom-

plished by describing the motion, not in ordinary space, but in the so-called phase space, in which there are dimensions associated with both the generalized coordinates and the generalized momenta. And the importance of Hamilton's equations arises from the fact that they are peculiarly suited to a discussion of motion by means of the phase space.

52. The Phase Space.—For a system with n degrees of freedom and n generalized coordinates $q_1 \dots q_n$, the phase space is a $2n$ -dimensional space in which $q_1 \dots q_n$ and $p_1 \dots p_n$ are plotted as variables. A single point in this space, often called a representative point, then determines all coordinates and velocities of the system. As time goes on, the representative point moves about the space, as both coordinates and momenta change with time. It is here that we make connection with Hamilton's equations, for these equations, $dq_i/dt = \partial H/\partial p_i$, $dp_i/dt = -\partial H/\partial q_i$, give just the components of the velocity of the representative point in the phase space. The problem of dynamics is to investigate the path of the representative point in the phase space.

We can easily see some properties of this motion. In the first place, it takes place with constant energy, assuming that we are dealing only with conservative forces. To prove this, we have for the time rate of change of H

$$\begin{aligned} \frac{dH}{dt} &= \frac{\partial H}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial H}{\partial q_2} \frac{dq_2}{dt} + \dots + \frac{\partial H}{\partial p_1} \frac{dp_1}{dt} + \frac{\partial H}{\partial p_2} \frac{dp_2}{dt} + \dots \\ &= \frac{\partial H}{\partial q_1} \frac{\partial H}{\partial p_1} + \frac{\partial H}{\partial q_2} \frac{\partial H}{\partial p_2} + \dots + \frac{\partial H}{\partial p_1} \left(-\frac{\partial H}{\partial q_1} \right) + \frac{\partial H}{\partial p_2} \left(-\frac{\partial H}{\partial q_2} \right) + \dots \\ &= 0. \end{aligned}$$

Now the energy H is a function of the coordinates and momenta, and hence a function of position in the phase space. Thus the equation $H = \text{constant}$ determines a single relation between all the p 's and q 's, and hence is the equation of a $(2n - 1)$ -dimensional hypersurface in the $2n$ -dimensional space. The representative point now moves about, but always stays on a single energy surface. If in addition there are other quantities which stay constant, as, for instance, an angular momentum, each one of these quantities gives an additional equation between the p 's and q 's, so that the representative point can move only in the intersection of all the various surfaces represented by these equations. Thus in some cases the region in which the motion

occurs is of smaller dimensionality than $2n - 1$. The extreme case is purely periodic motion; in that case there are enough quantities staying constant so that the motion of the representative point is in a single closed line in phase space, or a one-dimensional region. This fits in with the fact that all one-dimensional conservative motions not extending to infinity are periodic: for these have $n = 1$, $2n - 1 = 1$, so that the energy "surface" itself reduces to a line. Motions are possible in all the intermediate cases between the periodic motion, and the other extreme in which the representative point comes eventually arbitrarily close to every point of the energy surface. The latter type of motion is called quasi-ergodic, (ergodic motion being a nonexistent type in which the point passes through every point of the surface). Some of the intermediate types are multiply periodic motions, like our doubly periodic motion in the central field. We shall investigate some of these typical cases by means of examples.

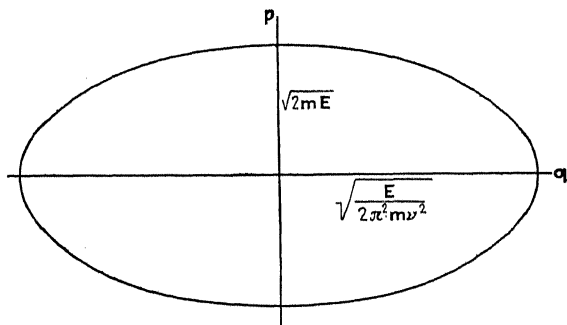


FIG. 11.—Phase space for a linear oscillator, with line of constant energy E .

53. Phase Space for the Linear Oscillator.—As an illustration of one-dimensional motion, we may take a linear oscillator (see Fig. 11). The phase space is two-dimensional, so that the energy surface is really a line. If the energy is $\frac{1}{2}mv^2 + 2\pi^2m\nu^2x^2$, where we readily see that ν is the frequency of oscillation, the Hamiltonian function is $p^2/2m + 2\pi^2m\nu^2x^2$. Setting this equal to a constant, E , the equation of the line of constant energy is $p^2/2m + 2\pi^2m\nu^2x^2 = E$, or

$$\frac{p^2}{(\sqrt{2mE})^2} + \frac{x^2}{(\sqrt{E/2\pi^2m\nu^2})^2} = 1, \quad (1)$$

the equation of an ellipse, having semi-axes $\sqrt{2mE}$ and $\sqrt{E/2\pi^2m\nu^2}$.

54. Phase Space for Central Motion.—As an illustration of a two-dimensional problem, we may take a central motion. The phase space is four-dimensional, so that we cannot directly plot it: the axes represent r , θ , p_r , p_θ . But we recall that in central motion p_θ stays constant, so that we may choose a particular value of p_θ , and use a three-dimensional section of the phase space, the axes representing r , θ , and p_r . We imagine r and θ as rectangular coordinates in a plane, and p_r as a coordinate at right angles to the plane. Now the energy surface is given by

$$\frac{p_r^2}{2m} + \frac{p_\theta^2}{2mr^2} + V(r) = E = \text{constant}, \quad (2)$$

or solving for p_r , $p_r = \pm \sqrt{2m(E - V - p_\theta^2/2mr^2)}$. That is, for each value of r and θ (E and p_θ being fixed), two values of p_r are given by the equation. If we plot these values, we get the surface of Fig. 12, on which the representative point moves in a spiral around the cylindrical surface, θ continually increasing, while r increases and decreases as the point spirals round and round. Although the orbit criss-crosses on itself, as we saw in

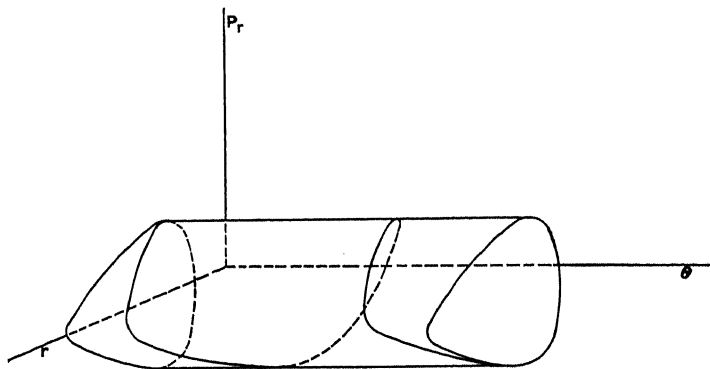


FIG. 12.—Surface of constant energy and constant angular momentum in phase space of a particle moving in a central field. The spiral represents the path of a particle.

Fig. 10, Chap. VII, still in the phase space the two different possible directions of motion at a given point of space are on opposite sides of the energy surface.

In Fig. 12, we have plotted only the part of the energy surface between $\theta = 0$ and $\theta = 2\pi$. The spiral, however, continues indefinitely. Since the regions from $\theta = 2\pi$ to 4π , 4π to 6π , etc., all represent the same regions of space as 0 to 2π , it is reasonable

to telescope these sections of the surface on to the one we have drawn. Each one will have its own segment of spiral, so that we shall have an infinite number of pieces all drawn on the surface shown in Fig. 12. There are now two possibilities. First, the motion can be periodic, as mentioned in Chap. VII. Then infinitely many segments of the spiral will lie on top of each other, resulting in one or a finite number of segments only. This is then a one-dimensional line in phase space, as we expect for periodic motion. Or second, the motion can be doubly periodic, the general case for this problem. In that case, the infinite number of segments of the spiral will not coincide, and instead they will fill the whole surface densely. In other words, in this case the path of the representative point fills a two-dimensional region. This is characteristic of doubly periodic motions.

55. Noncentral Two-dimensional Motion.—Let us consider motion in a field only slightly different from a central one, as, for instance, if we had a central field and a small external field of some other sort superposed. Then there will be slight torques acting on the particle, so that its angular momentum will change slowly. Now a given angular momentum corresponds to a given surface in Fig. 12. Hence in this motion the representative point does not confine itself to the surface we have drawn, but moves also on larger and smaller surfaces. If the motion without the additional torques were doubly periodic, and if no new regularities were introduced, the path of the representative point would now fill densely a whole set of surfaces with continuously varying sizes; that is, it would fill up a three-dimensional volume, the most general thing possible. In most cases this volume would be the whole region consistent with the energy, so that the motion would be quasi-ergodic. The motion itself, in two-dimensional coordinate space, would resemble for a short time the orbit of Fig. 10, Chap. VII, but the circles to which the orbit is tangent would slowly increase or decrease in size, the loops of the orbit simultaneously getting less or more rounded. If the departure from a central field were large, we could not use this approximate description, but should have to say simply that successive loops of the orbit were not merely oriented differently, but were of different size and shape.

56. Configuration Space and Momentum Space.—It is not always so easy to reduce a four-dimensional phase space to three dimensions as it was with the central field. We can always

visualize the phase space, however, by imagining a separate n -dimensional momentum space associated with each point of the n -dimensional coordinate or configuration space. If we assume that the n coordinates are the rectangular or Cartesian coordinates, then we have a simple interpretation of the condition that the representative point move on an energy surface. For this states that kinetic energy $= E - V$, or $(p_x^2 + p_y^2 + p_z^2) = 2m(E - V)$. But $p_x^2 + p_y^2 + p_z^2$ is simply the square of the radius in momentum space, so that to a given energy and a given point of space corresponds a sphere (or, with two dimensions, a circle) in the momentum space, on the surface of which the representative point must move. In quasi-ergodic motion, the representative point at one time or another comes arbitrarily close to each point of the surface of these spheres. We note that the spheres exist, and have real radii, only in that part of configuration space where $E - V$ is positive, and where, therefore, according to the energy principle, the motion can occur. But now in the more specialized types of motion, all points of the surface of the spheres are not available for representative points. Thus, in central motion, where the sphere degenerates to a circle in the two-dimensional momentum space, only those velocities are allowed which correspond to a given angular momentum. That is, p_x and p_y must satisfy at the same time the equations $p_x^2 + p_y^2 = 2m(E - V)$, $xp_y - yp_x = \text{angular momentum} = \text{constant}$, the equations of a circle and a straight line, respectively, in momentum space. These intersect in two points or in no points, so that for some parts of the configuration space corresponding to positive kinetic energy there are two possible values of the momentum, and for other parts there is none, and the motion cannot occur. These excluded regions are those within the small circle in Fig. 10, Chap. VII, and outside the large circle, but within the circle on which the kinetic energy becomes zero.

57. The Two-dimensional Oscillator.—A second example of two-dimensional motion is provided by the two-dimensional oscillator. In Chap. VII, Probs. 7, 8, and 9, it was shown that a particle attracted to a center by the forces $F_x = -ax$, $F_y = -ay$, could be solved by separation of variables, each coordinate vibrating like a separate oscillator, and the combined motions producing an elliptical orbit with the center at the center of attraction. The motion is periodic, with the same period which the corresponding one-dimensional motion would have. To

obtain a nonperiodic motion, we make $F_x = -cx$, $F_y = -ky$, the force components being proportional to the displacements, but with different coefficients. It is easily seen that in this case the total force, regarded as a vector, is not in the same direction as the displacement. An example is found in the vibration of a rectangular stick, if one end is clamped and the other vibrates, since the stick is stiffer for bending in one direction than the

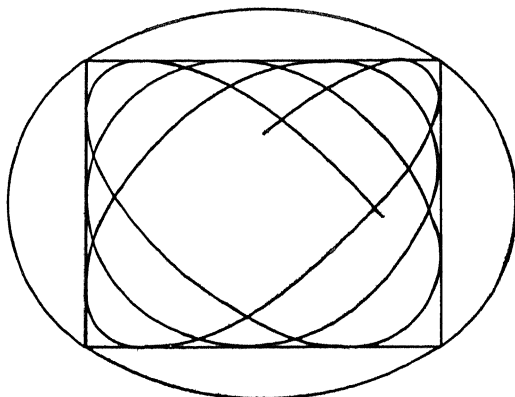


FIG. 13.—Lissajous figure for the orbit of a two-dimensional oscillator. The ellipse surrounding the rectangle represents the equipotential corresponding to the energy of the motion.

other, unless it is square. The variables are still separated in the equations of motion, and the solution is

$$x = A_1 \cos (\sqrt{c/m} t - \alpha_1), y = A_2 \cos (\sqrt{k/m} t - \alpha_2). \quad (3)$$

The motion is no longer periodic, for after one period of the x motion, the y motion will not have traversed just a full period, but will be in a different phase. By plotting (see Fig. 13), one can see that the orbit is always within the rectangle bounded by $x = \pm A_1$, $y = \pm A_2$, that it is often tangent to the edges of this rectangle, and that in time it comes arbitrarily close to any point within the rectangle. The resulting figure is called a Lissajous figure, and this sort of motion is typical of many examples which one meets. The orbit in central motion is, in fact, a sort of Lissajous figure, as Fig. 10, Chap. VII, shows.

The two-dimensional oscillator is a typical doubly periodic motion, the periods being just those of the two separate degrees of freedom. The displacements x and y are singly periodic, but if we wished to express, for example, the displacement in an arbitrary direction as a function of time, we should have $ax + by$,

which would be a sum of two terms, one periodic with the one frequency, the other with the other. Inspection of Fig. 13 shows that at a given point of space, there are just two branches of the orbit, corresponding to two definite values of the momentum, rather than having all momenta consistent with the given kinetic energy, in all directions, permitted as in quasi-ergodic motion.

A small perturbation applied to the two-dimensional oscillator would destroy the double periodicity, and make the motion quasi-ergodic. Thus we might have a small central field added to the linear restoring force. If the perturbation were small, we might apply what is called the method of variation of constants. That is, we could consider the coordinates to be given by Eq. (3), but regard the A 's and α 's as slowly varying functions of time rather than constants. Substituting these expressions in the differential equations, we should find that the perturbation produced such changes of amplitude and phase, at rates proportional to the magnitude of the perturbation. Considered from the standpoint of Fig. 13, this means that the rectangle is gradually changing its dimensions, subject always, however, to the condition that it is at least approximately inscribed in the same ellipse, since the total energy is only slightly changed by the perturbation. The result then is a slowly changing Lissajous figure, looking therefore like a superposition of many such figures, filling up the ellipse, and giving at a point of space not two possible momenta only, but a continuous range of momenta, in all directions, leading, therefore, to quasi-ergodic motion. A similar discussion can be given for the simpler problem of the almost periodic oscillator. In the exactly periodic case, the orbit is a single ellipse inscribed in a rectangle like that of Fig. 13, which in turn is inscribed in an ellipse. If the problem is made slightly different, by introducing only a very small difference between the force constants in the two directions, the dimensions of the ellipse can be considered to change slowly, though it always remains inscribed in the rectangle. The actual Lissajous figure, as one sees at once by inspection, is very similar to what one would obtain by drawing a great many ellipses, all inscribed in the same rectangle.

58. Methods of Solution.—We have seen one method of solving mechanical problems in several dimensions, that of separation of variables, by which the problem is reduced essentially to

several independent one-dimensional problems. There are several problems which can be solved by this method, in addition to the oscillator and the central field problems which we have treated. The problem of a particle in the field of two attracting centers, both attracting according to the inverse square law, can be solved by separation in ellipsoidal coordinates, with the two centers as foci. In three dimensions, the central or the axially symmetrical fields can be solved by separation. The solutions in all these cases are multiply periodic, as we can see at once from the fact that each coordinate, acting like a one-dimensional problem, must be singly periodic. It is thus obvious that no problems except multiply periodic ones can be solved by separation, and it seems likely that the small list we have just given includes practically all the multiply periodic mechanical problems which exist in two or three dimensions.

59. Contact Transformations and Angle Variables.—Hamilton's equations can be applied to multiply periodic motions by making certain transformations of coordinates which are called contact transformations, because it can be shown that they transform two curves which are in contact with each other in the original space into curves in contact in the new space. An ordinary transformation of coordinates, of the sort which we have discussed in connection with Lagrange's equations, is a transformation in which new coordinates are written as functions of the old ones: $q_i' = q_i'(q_1 \cdots q_n)$, if the q 's are the old coordinates, the q' 's the new ones. The new momenta, derived from the new Lagrangian function, are then functions of the old coordinates and momenta: $p_i' = p_i'(q_1 \cdots q_n, p_1 \cdots p_n)$. Such a transformation is called a point transformation. But in a contact transformation, the new coordinates as well as the new momenta are functions of both the old coordinates and momenta:

$$\begin{aligned} q_i' &= q_i'(q_1 \cdots q_n, p_1 \cdots p_n), \\ p_i' &= p_i'(q_1 \cdots q_n, p_1 \cdots p_n). \end{aligned} \quad (4)$$

There must naturally be restrictions on the functions, just as in ordinary point transformations we require that the new momenta be derived from the new Lagrangian function. When these restrictions are applied, however, it proves that Hamilton's equations are still satisfied in the new coordinates, though Lagrange's are not. Such contact transformations can often

be very useful in complicated problems, reducing them to forms which can be handled mathematically. A contact transformation can be most easily visualized simply as a change of variables in the phase space. For instance, suppose we have the phase space for a linear oscillator, as in Fig. 11. We can easily choose the scale so that the line of constant energy is a circle, rather than an ellipse. Then it is often useful to introduce polar coordinates in the phase space, so that the motion is represented by a constant value of r , and a value of θ increasing uniformly with time. The angle θ , or rather $\theta/2\pi$, in this case, is often called the angle variable, and is used as the coordinate. This is from analogy with the rotation of a body acted on by no torques, where the angular momentum stays constant, and the angle increases linearly with the time. The momentum conjugate to the angle variable, which stays constant with time, is not simply the radius, as we should expect from the simple use of polar coordinates, but proves to be proportional to the square of r ; in fact, it is just πr^2 , or the area of the circle. This momentum is called the action variable, or phase integral, denoted by J , and the angle variable is denoted by w .

Since Hamilton's equations hold in the transformed coordinates, and since evidently the energy H depends only on J , being independent of w , Hamilton's equations become

$$-\frac{\partial H}{\partial w} = 0 = \frac{dJ}{dt}, \quad (5)$$

verifying the fact that J is a constant of the motion; and

$$\frac{\partial H}{\partial J} = \frac{dw}{dt}, \quad (6)$$

a quantity independent of time, and of w , verifying the fact that w increases uniformly with time. Now since $w = \theta/2\pi$, it increases by unity in one period, so that dw/dt is just $1/T$, where T is the period, or is ν , the frequency of motion. Hence we have the important relation that

$$\nu = \frac{\partial H}{\partial J}, \quad (7)$$

giving the frequency of motion in terms of the derivative of the energy with respect to the action variable J .

It can be shown in a similar way that action and angle variables can be introduced in general in one-dimensional periodic motions. In every case the w 's increase uniformly with time, the frequency being given by Eq. (7). It also proves to be true in general that the action variable J is given by the area of the path of the representative point in phase space, which is the reason why it is called a phase integral. This area can be written $\oint p \, dq$, where this is analogous to $\int y \, dx$, the area under the curve $y(x)$. In Fig. 11, for instance, we integrate from the minimum to the maximum q along the upper branch of the ellipse, obtaining the part of the area above the q axis; then integrate back along the lower branch, where both p and dq are negative, obtaining the area below the q axis, so that the complete integral about the whole curve, which may be written $\oint p \, dq$, gives the whole area, or J . Connected with this is the criterion which a transformation of the p 's and q 's must satisfy if it is to be a contact transformation: it can be proved that it is a transformation in which areas in the phase space are preserved, or are not affected by the transformation, though the shape of an area in the new coordinates may be very different from what it was in the old. An immediate result of this is that the J 's are the same no matter what coordinates we may use for computing them.

Angle variables can also be introduced in cases with several degrees of freedom, provided the motion is multiply periodic, by using a separate angle variable for each coordinate. It is evident that the method could not be used with motions which were not multiply periodic, for we have seen that it is only in the multiply periodic motions that there are quantities, as, for example, angular momenta, which stay constant. Yet the action variables, or J 's, must stay constant, and consequently cannot be introduced, for example, in quasi-ergodic motions, where by hypothesis constants of the motion of this sort do not exist.

We shall meet angle variables and phase integrals again in Chap. XXX, where it is seen that they have close connection with the quantum theory. In that theory, the phase integrals prove to be quantized; that is, they take on only discrete values, J being limited to integral multiples of a fundamental physical constant, Planck's constant h ; and Eq. (7) for frequencies is replaced by an equation of finite differences, ν being a difference of energy in two energy levels, divided by the corresponding

difference of J (which can be simply h). These two formulas, which we elaborate later, form the basis of much of quantum theory.

60. Methods of Solution for Nonperiodic Motions.—When we meet a problem whose solution is quasi-ergodic, we are facing a branch of mathematics which offers no explicit or exact solutions. The only solutions are in the form of various series methods, for instance by the method of perturbations, which can be used if the motion is almost multiply periodic. We indicated an example of this in discussing the two-dimensional oscillator, where we treated the problem as a Lissajous figure with slowly varying amplitudes and phases. In general, the method of perturbations consists in developing the various quantities which appear in the problem in power series in the small quantities measuring the deviation from the multiply periodic case. If, for instance, that case has been discussed by the method of angle variables, we regard the J 's as slowly varying functions of time, their rate of variation being proportional to the first order to the magnitude of the perturbation. But in all these methods there is great difficulty in the matter of the convergence of the series; as time goes on, or as we consider larger and larger perturbations, they converge worse and worse, as is natural from the physical fact that often a slight change in initial conditions may, after the lapse of enough time, cause a profound change in the motion. These difficulties, as well as these methods of solution, are met particularly in celestial mechanics.

Problems

1. Given a linear oscillator of mass m , frequency ν , displacement x , momentum p , we can introduce a new coordinate w and momentum J , by the transformation

$$\begin{aligned}x &= \sqrt{J/2\pi^2 m \nu} \cos 2\pi w \\p &= -\sqrt{2mJ\nu} \sin 2\pi w.\end{aligned}$$

This change of variables can be shown to be a contact transformation. Find the Hamiltonian in terms of the new variables, by substituting these values of x and p in the total energy. Show that this resulting Hamiltonian depends on J alone, being independent of w , and show that w is an angle variable. Verify that J is the phase integral, or area enclosed by the orbit in the phase space, and that $\nu = \partial H/\partial J$. Show the geometrical interpretation of the contact transformation in the phase space.

2. An electron of charge $-e$, mass m , moves about a nucleus of charge Ze , and very large mass. The potential energy is $-Ze^2/r$. Assuming the energy to be E , angular momentum p_θ , separate variables, and consider

the radial motion as a one-dimensional problem, as in Chap. VII. Take a two-dimensional phase space in which r and p_r are variables, and plot the path of the representative point in this space.

3. Find the area of the path of the representative point in Prob. 2, and show that it is $\sqrt{2\pi^2 m Z^2 e^4 / (-E)} - 2\pi p_\theta$. Set this equal to J_r , the action variable connected with the radial motion. Find the energy in terms of J_r , and by differentiation find the frequency of motion. Verify this result in the special case of circular motion, where you can compute the rotational frequency by elementary methods.

4. If $F_x = -cx$, $F_y = -ky$, prove by direct calculation that the force, regarded as a vector, is at right angles to the equipotential. Show that the force is not in the direction of the displacement.

5. Suppose in a two-dimensional oscillator that the force constants along the two axes are only slightly different from each other. Prove that the orbit resembles an ellipse, of slowly changing shape and size. (Hint: show that $x = A \cos(\omega t - \alpha)$, $y = B \cos(\omega t - \beta)$, where A , B , α , and β are constants, is the equation of the ellipse. Then show that the equation of the path of the oscillator can be written in this form, if α and β are slowly changing functions of time.)

6. A particle moves as if it were executing simple harmonic motion about the center of a turntable, and at the same time the turntable were rotating with uniform angular velocity. Compute the x coordinate of the particle as a function of time, and show that the motion is doubly periodic.

7. Sketch the orbits in Prob. 6, for several different ratios between the frequencies of oscillation and rotation, including some cases of irrational ratios, and also simple rational ratios, as $1/1$, $1/2$, $2/1$.

8. A particle moving in two dimensions is attracted by two centers, of the same strength, attracting with a force proportional to the inverse square of the distance. Compute and plot a number of equipotentials, showing that for some energies the motion must be entirely confined to the region around one or the other center, while for larger energies it can surround both centers.

9. A particle moves in three dimensions under the action of a force of attraction to a center, depending only on the distance. Set up the problem in spherical coordinates, using the results of Probs. 3 and 4, Chap. VIII. Show that the variables can be separated, so that the problem is multiply periodic. Show that energy, total angular momentum, and the component of angular momentum along the axis of coordinates, all remain constant, showing the connection of these quantities with the generalized momenta of the problem. Using the obvious fact that the motion occurs in a plane and is just like two-dimensional central motion in that plane, show that the periods of the motions in θ and ϕ are the same, so that the motion is only doubly, not triply, periodic.

CHAPTER X

THE MOTION OF RIGID BODIES

In the preceding chapters we have been treating the mechanics of particles. Then we have passed on to the general methods of Lagrange and Hamilton, which can be applied to all sorts of mechanical problems. The present chapter will take up the motion of rigid bodies.

In elementary work, one learns the main outlines of the problem of the motion of a rigid body. We know that its motion is a superposition of a translation and a rotation. There are two fundamental laws of motion: the force equals the time rate of change of linear momentum, and the torque equals the time rate of change of angular momentum. To make our ideas more precise, the translational motion generally refers to the motion of the center of gravity, and the rotational to rotation about the center of gravity. The motion of the center of gravity is essentially like the motion of a particle, which we have already treated. In order to leave that out in the present chapter, we shall assume that no net forces act, or that the body is pivoted, rotating about a fixed point.

61. Elementary Theory of Precessing Top.—A torque is a vector, equal in magnitude to the force acting times its lever arm (that is, the perpendicular distance from the center of rotation to the line of action of the force), and at right angles to force and lever arm. That is, in vector notation, the torque on a single particle is $(\mathbf{r} \times \mathbf{F})$, where \mathbf{r} is the radius vector to the particle, \mathbf{F} the force acting, and the torque on the whole body is the vector sum of the separate torques on its parts. Similarly the angular momentum is a vector, defined in an analogous way: the angular momentum of a particle is equal in magnitude to the momentum times its lever arm, and at right angles to both, so that it is $[\mathbf{r} \times (m\mathbf{v})]$, or $m(\mathbf{r} \times \mathbf{v})$, and the total angular momentum of the body is the vector sum of the angular momenta of its parts. We see then that the equation "torque equals time rate of change of angular momentum" is a vector equation. This results in having two separate sorts of effect which a torque can produce. For we can analyze the torque into two components, one parallel

to the angular momentum, the other at right angles. The first component of torque produces an increase or decrease of angular momentum in the same direction as the angular momentum already existing; that is, it produces a speeding up or slowing down of the rotation, or an ordinary angular acceleration. This is the effect seen in the speeding up or slowing down of wheels on fixed axes. The component of torque at right angles to the angular momentum, on the other hand, produces a rotation or precession of the angular momentum vector, without change of length, and hence a change in the axis of rotation. This is the effect considered in the simple theory of the symmetrical top: if p represents the angular momentum of the top at a given instant (see Fig. 14), which is in approximately the same direction as the axis of figure of the top, the torque of gravity on the top will be $mg l$

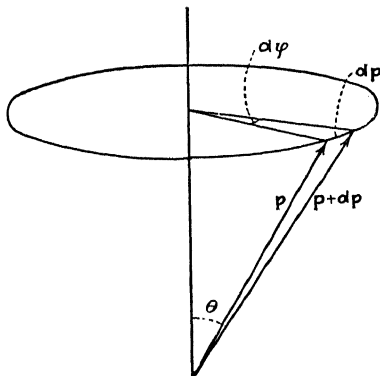


FIG. 14.—Angular momentum vectors for precessing top. The increment of angular momentum dp , proportional to and in the same direction as the torque of gravity, changes the total angular momentum from p to $p + dp$, resulting in a precession through the angle $d\phi$.

in magnitude, where l is the distance from the point of support to the center of gravity. The torque will act at right angles to the axis and p , so that the change of momentum in time dt will be dp , as shown. Thus the angular momentum after time dt will be the vector $p + dp$, obtained from the old vector by a precession, as if the whole figure were rotated about the axis through the angle $d\phi$. We can easily find the rate of precession. For $d\phi$ evidently equals dp divided by the

radius of the circle, or is $\frac{dp}{|p| \sin \theta}$. On the other hand, $dp = mg l \sin \theta dt$. Hence $\frac{mg l \sin \theta dt}{|p| \sin \theta} = d\phi$, or $\frac{d\phi}{dt} = \frac{mg l}{|p|}$, a precession

increasing with increasing torque, but decreasing with increasing angular momentum. We note that if we regard the precessional velocity as a vector, say ω , along the vertical direction, and having

a magnitude $\frac{mg l}{|p|}$, we have

$$\frac{dp}{dt} = (\omega \times p). \quad (1)$$

This is a general relation for a precessing vector, as we readily see.

The elementary ideas of torque and angular momentum do **not** permit us to go much farther than we have indicated here, without further analysis. With a body in the absence of torques, for instance, we know at once that the angular momentum stays constant, both in direction and in magnitude. But this tells us little about the actual complicated motion. We must then examine the problem more in detail. In the succeeding sections we consider the angular momentum, kinetic energy, etc., of solid bodies of arbitrary shape, with arbitrary axes of rotation, though we always assume that they rotate about a fixed point, as the center of gravity.

62. Angular Momentum, Moment of Inertia, and Kinetic Energy.—Let a body rotate about the origin as a center, the axis of rotation having direction cosines λ , μ , ν with the three axes. We may regard the angular velocity as a vector, whose direction is the axis of rotation, and whose magnitude is the magnitude of angular velocity. Thus, if the vector is ω , its magnitude ω_0 , we have $\omega_x = \lambda\omega_0$, $\omega_y = \mu\omega_0$, $\omega_z = \nu\omega_0$. Now we can easily find the linear velocity of any point of the body. This is numerically $\rho\omega_0$, where ρ is the perpendicular distance from the point to the axis of rotation, and is at right angles to the axis of rotation and the perpendicular distance. In other words, the velocity is given by the vector product $(\omega \times r)$; a little consideration shows that the vector product has the right direction. Now that we know the velocity of each point, we can compute the angular momentum. We have already seen that this is the sum of terms $m(r \times v)$ for all particles of the body. But $v = \omega \times r$, so that angular momentum = $\Sigma m[r \times (\omega \times r)]$. This can be easily expanded. Thus the x component, for example, is

$$\begin{aligned} & m[y(\omega \times r)_z - z(\omega \times r)_y] \\ &= m[y(\omega_x y - \omega_y x) - z(\omega_z x - \omega_x z)] \\ &= m\omega_x(y^2 + z^2) - m\omega_y xy - m\omega_z xz, \end{aligned}$$

with corresponding formulas for the other components. If now we sum over all particles of the body, remembering that ω is the same for all, we have, if p_x , p_y , p_z are the components of angular momentum,

$$\begin{aligned}
 p_x &= A\omega_x - F\omega_y - E\omega_z, \\
 p_y &= -F\omega_x + B\omega_y - D\omega_z, \\
 p_z &= -E\omega_x - D\omega_y + C\omega_z,
 \end{aligned} \tag{2}$$

where for abbreviation we set $A = \Sigma m(y^2 + z^2)$, $B = \Sigma m(z^2 + x^2)$, $C = \Sigma m(x^2 + y^2)$, $D = \Sigma myz$, $E = \Sigma mzx$, $F = \Sigma mxy$. The quantities A , B , and C are called the moments of inertia, and D , E , F are the products of inertia; the first three are obviously the moments of inertia of the body in the ordinary sense, about the x , y , and z axes, respectively. We note one thing at the outset: the angular momentum vector is not in general parallel to the angular velocity vector. Thus if $\omega_y = \omega_z = 0$, so that the angular velocity is along the x axis, we have all three components of p in general different from zero.

Next we find the kinetic energy. For a single particle, this is $\frac{1}{2}mv^2$, or $\frac{1}{2}m(\omega \times r)^2$. Again expanding, this is

$$\begin{aligned}
 \frac{1}{2}m[(\omega_y z - \omega_z y)^2 + (\omega_z x - \omega_x z)^2 + (\omega_x y - \omega_y x)^2] \\
 = \frac{1}{2}m[\omega_x^2(y^2 + z^2) + \omega_y^2(z^2 + x^2) + \omega_z^2(x^2 + y^2) \\
 - 2\omega_x\omega_y xy - 2\omega_y\omega_z yz - 2\omega_z\omega_x zx].
 \end{aligned}$$

Summing over all particles, and using the abbreviations above, this is

$$T = \frac{1}{2}(A\omega_x^2 + B\omega_y^2 + C\omega_z^2 - 2D\omega_y\omega_z - 2E\omega_z\omega_x - 2F\omega_x\omega_y). \tag{3}$$

The quantity T can be written as $\frac{1}{2}I\omega_0^2$, where

$$I = \lambda^2 A + \mu^2 B + \nu^2 C - 2\mu\nu D - 2\nu\lambda E - 2\lambda\mu F. \tag{4}$$

It is easily shown that I is simply $\Sigma m\rho^2$, where ρ , as before, is the perpendicular distance from the point to the axis of rotation, so that I agrees with the elementary definition. If we imagine λ , μ , and ν varied in any manner, the quantities A , B , . . . F do not change. As a variation in λ , μ , ν means a variation of direction of the rotation axis through the center of rotation O , we see that the sums A . . . F completely determine the moment of inertia of the body about any axis through the same center of rotation.

63. The Ellipsoid of Inertia; Principal Axes of Inertia.—The Eq. (4) for the moment of inertia I may be interpreted geometrically in a very simple manner. The equation $Ax^2 + By^2 +$

$Cz^2 - 2Dyz - 2Ezx - 2Fxy = \text{constant}$ represents a surface of second degree. If we denote by r the radius vector drawn from O to a point on this surface, having direction cosines λ, μ, ν , Eq. (4) becomes

$$r^2(A\lambda^2 + B\mu^2 + C\nu^2 - 2D\mu\nu - 2E\nu\lambda - 2F\lambda\mu) = \text{constant}. \quad (5)$$

The expression inside the parentheses is just I , the moment of inertia, so that we have $r^2 = \text{constant}/I$, or $I = \text{constant}/r^2$. Now, since the moment of inertia is always positive and can never vanish, r^2 cannot become infinite and our surface is a closed surface. Since it is of second degree it is an ellipsoid with its center at O , and is called the ellipsoid of inertia at the point O . The ellipsoid of inertia has the simple physical significance that the moment of inertia of the body about any axis through O is measured by the inverse square of the radius vector from O , drawn parallel to the rotation axis and terminating on the surface of the ellipsoid.

Every ellipsoid has three principal axes which are mutually orthogonal. These axes are known as the principal axes of inertia at O . Just as in the case of an ellipse, when coordinate axes are chosen coincident with the principal axes, the equation of the ellipsoid reduces to a sum of squares, so that the coefficients of the terms in yz , zx , and xy disappear and we have $D = E = F = 0$. We shall often use coordinate axes coincident with the principal axes, but since these axes are fixed with respect to the rigid body, we must always remember that they are rotating axes in space, and we must describe their motion with respect to a system of axes fixed in space. Referred to the principal axes, the moment of inertia becomes simply $\lambda^2 A_0 + \mu^2 B_0 + \nu^2 C_0$, where these A_0, B_0 , and C_0 are now computed with respect to axes fixed in the body, and so do not change with rotation of the body, as do the ordinary moments and products of inertia computed with respect to fixed axes. The kinetic energy of rotation is then $T = \frac{1}{2}I\omega_0^2 = \frac{1}{2}(\lambda^2\omega_0^2 A_0 + \mu^2\omega_0^2 B_0 + \nu^2\omega_0^2 C_0)$, which is also $T = \frac{1}{2}(A_0\omega_1^2 + B_0\omega_2^2 + C_0\omega_3^2)$, where ω_1, ω_2 , and ω_3 are the components of ω taken about the principal axes.

64. The Equations of Motion.—Suppose the moment, or torque, of the external force is M , with components M_x, M_y, M_z . Then the equations of motion are obtained by setting the torque equal to the time rate of change of angular momentum: $M = dp/dt$, or for the x component

$$M_x = \frac{d}{dt}(A\omega_x - F\omega_y - E\omega_z), \quad (6)$$

where, of course, we are using arbitrary x, y, z coordinates, not the principal axes. In performing the differentiation, we must remember that not only are $\omega_x, \omega_y, \omega_z$ changing, but also A, F, E , since the body is rotating, and these moments and products of inertia are defined with respect to a particular fixed coordinate system. Thus we have

$$M_x = A\dot{\omega}_x - F\dot{\omega}_y - E\dot{\omega}_z + \dot{A}\omega_x - \dot{F}\omega_y - \dot{E}\omega_z.$$

The last three terms can be rewritten, using, for instance,

$$\dot{A} = \frac{d}{dt} \sum m(y^2 + z^2) = 2 \sum m(y\dot{y} + z\dot{z}),$$

along with $v = \omega \times r$, so that $\dot{x} = \omega_y z - \omega_z y$, etc. Without trouble we find that the equations can be written

$$M_x = A\dot{\omega}_x - F\dot{\omega}_y - E\dot{\omega}_z - (B - C)\omega_y\omega_z - D(\omega_y^2 - \omega_z^2) + F\omega_x\omega_z - E\omega_x\omega_y, \quad (7)$$

with equivalent equations for the y and z components. The latter terms seem very complicated; but we readily see that they can be written as a vector product, giving

$$M_x = \frac{dp_x}{dt} = A\dot{\omega}_x - F\dot{\omega}_y - E\dot{\omega}_z + (\omega \times p)_x \quad (8)$$

The equation for time rate of change of angular momentum, in the form above, has a simple interpretation. Suppose we have any vector G , and that we consider it with respect to rotating coordinates, rotating with the angular velocity ω . If we were rotating with the coordinates, the vector would seem to have a certain time rate of change, which we may call $\partial G / \partial t$. But this will not be its actual time rate of change, when looked at from a stationary system of coordinates. For even a vector which remained constant in the rotating system would actually be changing, just on account of its rotation. In fact, the rate of change of the vector for this latter reason, using the same sort of argument which we met in Eq. (1) in describing the precessing top, is $(\omega \times G)$, and the total rate of change of G is the sum of these two effects, or

$$\frac{dG}{dt} = \frac{\partial G}{\partial t} + (\omega \times G). \quad (9)$$

In particular, then, with the angular momentum, we evidently have two terms of the sort considered above. We conclude therefore that

$$A\dot{\omega}_x - F\dot{\omega}_y - E\dot{\omega}_z = \left(\frac{\partial p}{\partial t} \right)_x,$$

so that these terms represent the rate of change of angular momentum, with respect to the rotating axes.

One result of the theorem we have just worked out is interesting. Let the vector G be the angular velocity. Then $d\omega/dt = \partial\omega/\partial t$, since the vector product $(\omega \times \omega)$ is zero. Hence the components of time rate of change of angular velocity are the same in fixed as in rotating axes.

65. Euler's Equations.—The equations of motion, (7) or (8), take on a particularly simple form when expressed in terms of the principal axes. Let us first take our fixed axes xyz so that they coincide with the instantaneous values of the rotating, principal axes. Then D, E, F are instantaneously zero, and the equations (7) are

$$M_x = A_0\dot{\omega}_x - (B_0 - C_0)\omega_y\omega_z,$$

with two similar equations. But now let $\omega_1, \omega_2, \omega_3$ be the components of angular velocity with respect to the rotating principal axes. Momentarily these equal $\omega_x, \omega_y, \omega_z$. But also $\dot{\omega}_1$ is the same thing as $(\partial\omega/\partial t)_x$, the x component of the time rate of change of angular velocity with respect to the rotating axes. We have just shown, however, that this equals $(d\omega/dt)_x$, or $\dot{\omega}_x$. Hence we can rewrite our equations entirely in terms of the moving axes,

$$\begin{aligned} M_1 &= A_0\dot{\omega}_1 - (B_0 - C_0)\omega_2\omega_3 \\ M_2 &= B_0\dot{\omega}_2 - (C_0 - A_0)\omega_3\omega_1 \\ M_3 &= C_0\dot{\omega}_3 - (A_0 - B_0)\omega_1\omega_2, \end{aligned} \tag{10}$$

where M_1, M_2, M_3 are the components of torque with respect to the rotating axes. These equations are called Euler's equations.

66. Torque-free Motion of a Symmetric Rigid Body.—We shall now apply Euler's equations to the motion of a rigid body symmetric about an axis, subject to the action of no external torques (either the external forces are zero, or act at the center of mass). The earth provides a good example, if we neglect the torques due to sun and moon. We choose the center of mass

as an origin, and take the axis of symmetry as principal axis 3. The principal moments of inertia are then A_0 , A_0 , C_0 . Euler's equations for this case are

$$\begin{aligned} A_0 \dot{\omega}_1 + \omega_2 \omega_3 (C_0 - A_0) &= 0 \\ A_0 \dot{\omega}_2 + \omega_1 \omega_3 (A_0 - C_0) &= 0 \\ C_0 \dot{\omega}_3 &= 0. \end{aligned}$$

The last equation integrates at once, giving $\omega_3 = \text{constant}$. This means that the resultant angular velocity has a constant component along the axis of symmetry. If we now place $\alpha = \frac{C_0 - A_0}{A_0}$, the two other equations are $\dot{\omega}_1 + \alpha \omega_2 = 0$ and

$\dot{\omega}_2 - \alpha \omega_1 = 0$. Differentiating the first of these, we find $\ddot{\omega}_1 + \alpha \dot{\omega}_2 = \ddot{\omega}_1 + \alpha^2 \omega_1 = 0$, which has as its solution $\omega_1 = a \cos(\alpha t + \epsilon)$, and putting this value of ω_1 in the second equation, we find $\omega_2 = a \sin(\alpha t + \epsilon)$, where a and ϵ are integration constants. From these equations we see that the resultant angular velocity $\omega = \sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2} = \sqrt{a^2 + \omega_3^2}$ is constant, and that the projection of ω on the plane perpendicular to the axis of symmetry and fixed in the body describes a circle of radius a with a period given by $\tau = \frac{2\pi}{\alpha} = \frac{2\pi}{\omega_3} \frac{A_0}{C_0 - A_0}$. In the case of

the earth, $\omega_3 = 2\pi$ per day, so that τ becomes $A_0/(C_0 - A_0)$ days, which is about 300 days and is known as the Euler period. This period is not observed, but there is one of 427 days known as the Chandler period giving rise to a variation of latitude. When the imperfect rigidity of the earth is taken into account, it is possible to identify these two as the same.

We can get an idea of the actual motion most clearly from a diagram. In Fig. 15, we show an oblate spheroid, to represent the symmetrical body. There is a circular conical hole Obd cut out surrounding the north pole a , and a fixed cone touching the inside of this hole, and centered on the line Oc . The motion is now as if one cone rolled on the other. We see at once that, since the axis Ob is instantaneously at rest, it is the instantaneous axis of rotation ω . As time goes on, this axis of rotation traces out the cone Obd with respect to the body, and at the same time traces out the cone Obe fixed in space. The axis of the fixed cone, Oc , is the direction of the constant total angular momentum vector. Other properties of the motion are discussed in a problem.

67. Euler's Angles.—If we wish information about the general motion of the top, we must introduce some set of coordinates capable of describing its position. So far, we have not had any set of coordinates at all. We have worked with angular velocities, and angular momenta, which were vectors, and all the equations came out very neatly and symmetrically in terms of them. But there is a peculiar thing about the three components of angular velocity: there are no corresponding angles to serve as coordinates. This is not true in plane motions. If a body rotates

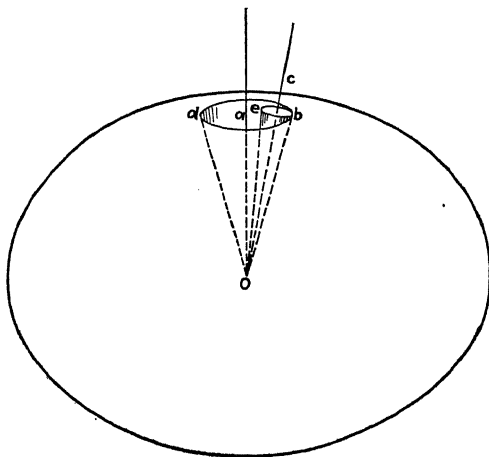


FIG. 15.—Space and body cones for the torque-free rotation of a symmetrical body. The cone Odb , fixed in the body, rolls on the cone Obe , fixed in space. The line Oa is the axis of symmetry of the body, Ob is the instantaneous axis of rotation, Oc the fixed axis of total angular momentum.

with angular velocity ω about a fixed axis, we can regard ω as $\dot{\theta}$, where θ is the angle through which the body has turned about the fixed axis, and which can be used as a coordinate. Then we can say that the component of angular momentum $I\omega$ is the momentum conjugate to θ , and the whole Lagrangian and Hamiltonian methods go through perfectly. As soon as we have three dimensions, however, and the possibility of different axes of rotation, we no longer have such angles. It is readily seen, for instance (we leave it for a problem), that one cannot use the angles through which the body has turned about the three coordinate axes as variables. The fact is that, though angular momentum is a vector, finite angular rotations are not, and do not have three components which can be used as coordinates.

We are forced, then, by the peculiar nature of angular rotations, to look for some set of three angles to describe the position of the body, which unfortunately cannot have the symmetrical nature of the x, y, z components of angular velocity. The usual set of angles are called Euler's angles, and are shown in Fig. 16. We ordinarily use these angles for discussing a symmetrical body. Then Oz is a fixed axis, for example, the vertical in the top problem. OC_0 is the axis of figure of the body, taken as the

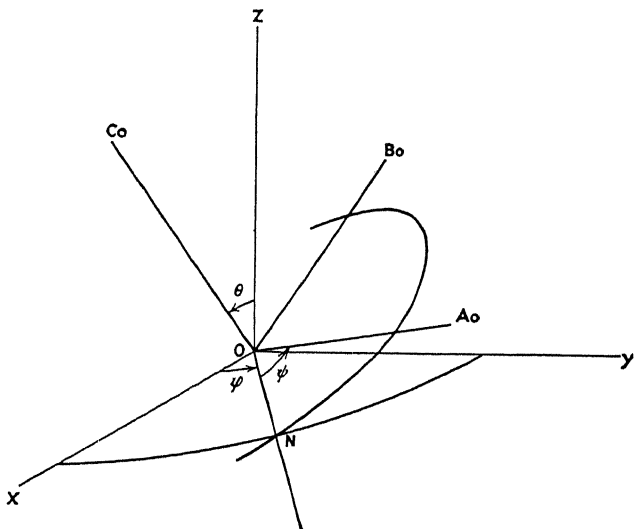


FIG. 16.—Euler's angles. For a symmetrical body, OC_0 is the axis of symmetry, OA_0 and OB_0 two axes fixed in the body at right angles. θ and ϕ measure colatitude and longitude of the direction of the principal axis; ψ measures the rotation of the body about the principal axis.

third principal axis. θ measures the angle between axis of figure and fixed axis, ϕ measures the angle of precession of the axis of figure, so that $d\phi/dt$ is the angular velocity of precession, and ψ measures the rotation of the body about its axis of figure measured from the line ON , called the nodal line. Thus we see that, though the Eulerian angles do not have symmetry, they are very natural ones for the problem in hand.

Let us set up the components of angular velocity, and the kinetic energy, in terms of the Eulerian angles. The motion of the body may be thought of as consisting of a rotation of the body about OC_0 and the motion of OC_0 relative to the fixed frame of reference. The former is described by the angular velocity $\dot{\psi}$ which has the components 0, 0, $\dot{\psi}$ (referred to the

principal axes). The latter motion consists of (a) a rotation $\dot{\theta}$ about the nodal line ON as an axis, which was zero in the steady motion of the top considered above; and (b) of a precession $\dot{\phi}$ about the z axis. The components of these angular velocities along the principal axes OA_0 , OB_0 , and OC_0 are

$$\begin{aligned} (a) \quad & \dot{\theta} \cos \psi; -\dot{\theta} \sin \psi; 0 \\ (b) \quad & \dot{\phi} \sin \theta \sin \psi; \dot{\phi} \sin \theta \cos \psi; \dot{\phi} \cos \theta. \end{aligned}$$

Adding these angular velocity components, we have

$$\begin{aligned} \omega_1 &= \dot{\theta} \cos \psi + \dot{\phi} \sin \theta \sin \psi \\ \omega_2 &= -\dot{\theta} \sin \psi + \dot{\phi} \sin \theta \cos \psi \\ \omega_3 &= \dot{\psi} + \dot{\phi} \cos \theta. \end{aligned}$$

Here $\dot{\psi}$ corresponds to the quantity ω_0 used for discussing the steady motion, and $\dot{\phi}$ to ω_1 . From these components of angular velocity, of course, we can at once get the angular momenta.

The kinetic energy, as we have seen, is $\frac{1}{2}(A_0\omega_1^2 + B_0\omega_2^2 + C_0\omega_3^2)$. But in our case of a symmetrical top this simplifies, since $A_0 = B_0$, and substituting we have

$$T = \frac{1}{2}[A_0(\omega_1^2 + \omega_2^2) + C_0\omega_3^2] = \frac{1}{2}[A_0(\dot{\theta}^2 + \sin^2 \theta \dot{\phi}^2) + C_0(\dot{\psi} + \dot{\phi} \cos \theta)^2]. \quad (11)$$

Using the kinetic energy, or corresponding Lagrangian function, in terms of the Eulerian angles, we can easily derive the Lagrangian equations of motion, and find them to be the same Euler equations which we have already obtained. For instance,

$$\text{using } L = T, \quad \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{\psi}} \right) - \frac{\partial L}{\partial \psi} = \frac{d}{dt} [C_0(\dot{\psi} + \dot{\phi} \cos \theta)] = C_0\dot{\omega}_3 = M_3,$$

which is the third of Euler's equations, when we remember that $A_0 - B_0 = 0$.

68. General Motion of a Symmetrical Top under Gravity.—

We are now ready to proceed with the general discussion of the top under gravity, for which we have already considered the steady precession. We note first that the torque is at right angles to the axis of figure. Hence by the third of Euler's equations, $\dot{\omega}_3 = 0$, or ω_3 is constant. Instead of using the other two of Euler's equations, it is somewhat more convenient to use the conservation of energy and of angular momentum to discuss the motion, much as we did in our earlier chapter on central motion. For the kinetic energy we have $T = \frac{1}{2}[A_0(\dot{\theta}^2 +$

$\sin^2\theta \dot{\phi}^2) + C_0\omega_3^2]$, and the potential energy is $Mgl \cos \theta$, where l is the distance from O to the center of mass. Thus the energy equation becomes

$$E = \frac{1}{2}(A_0(\dot{\theta}^2 + \sin^2\theta \dot{\phi}^2) + C_0\omega_3^2) + Mgl \cos \theta, \quad (12)$$

where E is the total energy. We now can eliminate ϕ from the equation above by utilizing the fact that there are no torques taken about the z axis. This means that the component of angular momentum along this axis is constant. The angular momentum due to the rotation ω_3 of the top about its axis of symmetry has a vertical component $C_0\omega_3 \cos \theta$. The angular velocity $\dot{\theta}$ of the axis contributes nothing to the vertical angular momentum. The other component of the angular velocity of the axis is $\sin \theta \dot{\phi}$, and this is about an axis perpendicular to OC , making an angle of $\pi/2 - \theta$ with the vertical. Thus the contribution of this term is $A_0 \sin^2\theta \dot{\phi}$, so that the conservation of angular momentum about the z axis yields

$$p_z = C_0\omega_3 \cos \theta + A_0 \sin^2\theta \dot{\phi}. \quad (13)$$

We now substitute the value of $\dot{\phi}$ taken from this equation into the energy equation and get a differential equation for θ alone, so that we may discuss the time variations of θ , or the variations of the inclinations of the axis of figure of the top with the vertical. When we make this substitution, and solve for $\dot{\theta}$, we have

$$\dot{\theta} = \sqrt{2(E - V')/A_0}, \quad (14)$$

where V' , which plays the part of a fictitious potential energy for the motion of this coordinate, has the value

$$V' = Mgl \cos \theta + \frac{1}{2}C_0\omega_3^2 + \frac{1}{2} \frac{(p_z - C_0\omega_3 \cos \theta)^2}{A_0 \sin^2 \theta}. \quad (15)$$

The first term is the gravitational energy, decreasing as the angle increases, showing that gravity tends to make the top fall. The second is a constant, the energy of the spinning motion. The third term is a dynamic term, reminding us of the centrifugal force term in the effective energy for the radial motion in a central field. It becomes infinite when $\theta = 0$ or π , since at those angles the rate of precession $\dot{\phi}$ would have to be infinitely rapid in order to conserve the angular momentum component p_z , contributing therefore an infinite amount to the energy. Between these angles this dynamic term has a single minimum. In other words, it exerts a stabilizing influence, quite apart from

any external forces which may act, and leads to a stable oscillation of θ about a certain minimum of V' , whose position is determined by the external torque.

69. Precession and Nutation.—The minimum of V' can be determined by differentiating with respect to θ , setting the result equal to zero. This gives

$$0 = -Mgl \sin \theta +$$

$$\frac{(p_z - C_0 \omega_3 \cos \theta)}{A_0 \sin^2 \theta} \left\{ C_0 \omega_3 \sin \theta - A_0 \cos \theta \sin \theta \left[\frac{p_z - C_0 \omega_3 \cos \theta}{A_0 \sin^2 \theta} \right] \right\}$$

or, from Eq. (13),

$$\phi[C_0 \psi - (A_0 - C_0)\phi \cos \theta] = Mgl. \quad (16)$$

If the energy is equal to the effective potential V' at this angle, $\dot{\theta}$ will be zero, and the motion is a pure precession of the sort described in Sec. 61. If we assume that the rate of precession

is small compared with the rate of rotation, which is the only case in which the angular momentum, the angular velocity, and the axis of figure are nearly enough in the same straight line so that the arguments of that section are valid, we have $\phi \ll \psi$. In that case the equation becomes $\phi(C_0 \psi) = Mgl$, $\phi = Mgl/C_0 \psi$, in agreement with the result of Sec. 61, when we recall that in this limit $C_0 \psi$ is approximately the total angular momentum. This condition, or rather the accurate condition (16), determines the rate of steady precession ϕ for any total angular momentum, a rate independent of θ to this approximation, but

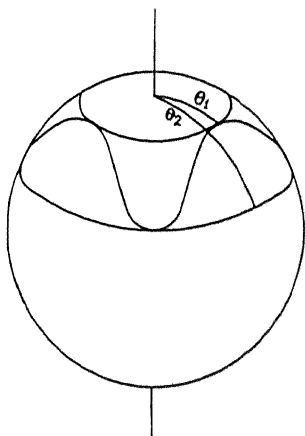


FIG. 17.—Nutation of a top. The sinusoidal curve is the projection of the axis of the top on a sphere. θ_1 and θ_2 are angular limits of the nutational motion.

depending on θ if we must consider terms in ϕ^2 .

If the energy E is greater than the minimum of V' , the curve of E will cut that for V' at two values of θ , one greater and one less than the inclination of the axis for the purely precessional motion which we have just discussed. In this case, θ will oscillate between these two limits. This oscillation is called nutation. The complete motion then consists of a combination of this nutation with a precession, as indicated in Fig. 17, where we draw

the intersection of the axis of the top with a sphere. The angles θ_1 and θ_2 are the two angles for which $E = V'$, so that the minimum of V' , or the angle for the pure precessional motion corresponding to the same angular momentum, lies between these two values. In the problems, the frequency of the nutational motion is discussed. We also discuss, in Prob. 9, the special case of the "sleeping" top, in which the top starts spinning vertically. In this special case, the dynamic term in V' is finite at $\theta = 0$, so that under certain circumstances oscillations about the vertical can occur.

Problems

1. Prove directly that the moment of inertia I , equal to $\Sigma m\rho^2$, is equal to $\lambda^2 A + \mu^2 B + \nu^2 C - 2\mu\nu D - 2\nu\lambda E - 2\lambda\mu F$, where λ , μ , ν , are the direction cosines of the axis of rotation.

2. Show that, if T is the kinetic energy of a rotating body, p its angular momentum, ω its angular velocity, $p_x = \partial T / \partial \omega_x$, and $2T = p_x \omega_x + p_y \omega_y + p_z \omega_z$.

3. In Fig. 15, show that $\tan \angle aOb = a/\omega_3$, and $\tan \angle aOc = \frac{A_0}{C_0} \frac{a}{\omega_3}$, where ω_3 , a represent the components of the angular velocity along and at right angles to the figure axis. Knowing that the time required for the axis Ob of angular velocity to perform a complete rotation with respect to the body is $\tau = \frac{2\pi}{\omega_3} \frac{A_0}{C_0 - A_0}$, show that the time for it to perform a complete rotation in

space is approximately $\frac{2\pi}{\omega_3} \frac{A_0}{C_0}$, if angles aOb and aOc are small. Hence show that for the earth the axis of angular velocity is not fixed, but rotates about a fixed direction approximately once a day.

4. The earth is acted on by torques exerted by the sun and moon, and as a consequence its angular momentum precesses about a fixed direction in space. This is entirely separate from the effect of Fig. 15 and Prob. 3, which we now neglect. This precession has a period of 25,800 years, and carries the angular momentum about a cone of semi-vertical angle $23^\circ 27'$, so that the pole in succession points to different parts of the heavens, resulting in the precession of the equinoxes, and in the fact that different stars act as pole star at different periods of history. Show that the motion can be represented by the rolling of a cone fixed in the earth, of diameter 21 in. at the north pole, on a cone of angle $23^\circ 27'$ fixed in the heavens.

5. A system of electrons moving about a center of attraction has a certain angular momentum, equal to $\Sigma m(r \times v)$, and also a magnetic moment, equal to $\Sigma \frac{e}{2c}(r \times v)$, where e is the charge and m the mass of an electron, c the velocity of light. This magnetic effect results because the electrons in rotation act like little currents, which in turn have magnetic fields like bar magnets. An external magnetic field H exerts a torque on the system, equal to the vector product of H and the magnetic moment. Show that

under the action of the field, the system of electrons precesses with angular velocity $e\hbar/2mc$ about the direction of the field. This precession, which, as we see, is independent of the velocities of the electrons, is called Larmor's precession.

6. One reason why finite rotations do not act as vectors is that they do not commute, that is, the same two rotations applied in one order lead to one answer, but in the opposite order to a very different answer. Demonstrate this by diagrams, imagining that we have a cube (label its faces by different letters or numbers on a diagram), originally in one position, with its edges parallel to the coordinate axes (position a). First rotate through 90 deg. about the x axis (position b), then through 90 deg. about the y axis (position c), drawing diagrams of each step. Then, starting again from position a , rotate first through 90 deg. about the y axis (position d) and then through 90 deg. about the x axis (position e). Show that (c) and (e) are entirely different orientations.

7. Write down the kinetic energy of a nonsymmetrical body, in terms of Euler's angles. Derive the Lagrangian equation for ψ , and show that it reduces to one of Euler's equations.

8. In the same way as in Prob. 7, set up the other two Lagrangian equations, showing that they lead to the other two of Euler's equations.

9. A top is started spinning vertically, with no other motion, so that initially $\theta = 0$, $d\theta/dt = 0$. Show that $p_z = C_0\omega_3$, $E = \frac{1}{2}C_0\omega_3^2 + Mgl$. Substituting these in the expression of Eq. (13) for $\dot{\theta}$, show that if $\omega_3 > \omega'$, where $(\omega')^2 = 4Mgl A_0/C_0^2$, the angle θ must remain equal to zero, but that if ω_3 falls below ω' , θ will oscillate between 0 and the angle $\cos^{-1} [2(\omega_3/\omega')^2 - 1]$. Experimentally, if a top is started as we have described, with $\omega_3 > \omega'$, there will be a frictional torque decreasing ω_3 , and as soon as the torque reduces ω_3 below ω' , the top will begin to wobble.

10. For a nutation of small amplitude about the steady precessional motion of a top, the angle θ oscillates sinusoidally about the equilibrium angle. Find the frequency of the nutation, by expanding the potential V' in power series in $\theta - \theta_0$, where θ_0 is the angle of steady precession with the same angular momentum. Retain only the constant and the term in $(\theta - \theta_0)^2$, and get the frequency by comparing with the corresponding expression for the linear oscillator.

CHAPTER XI

COUPLED SYSTEMS AND NORMAL COORDINATES

The mechanical problems which we have treated so far have been those where just one particle moved around, sometimes in a potential field, sometimes subject to forces not derivable from a potential. In many problems, however, there are several particles exerting forces on each other and influencing each other's motion. As examples, we have the actual solar system, where the sun, planets, and moons all act on each other; an atom, with the various electrons reacting; a molecule, with the atoms vibrating under the action of their mutual forces. A more familiar case is that of several electric circuits coupled together by induction or some other method. Another is that in which several pendulums or springs can react on each other, as through their supports, and affect each other's motion. There is evidently a very wide variety of problems; we shall treat only the simplest, in which two linear oscillators, or electric circuits, are coupled together by a force depending linearly on both the displacements.

70. Coupled Oscillators.—Suppose we have two undamped one-dimensional oscillators, whose displacements are y_1 and y_2 , respectively, and whose equations of motion, if uncoupled, would be

$$m_1 \frac{d^2 y_1}{dt^2} + k_1 y_1 = 0,$$
$$m_2 \frac{d^2 y_2}{dt^2} + k_2 y_2 = 0.$$

Now let them be acted on by equal and opposite forces proportional to the distance apart, $-a(y_1 - y_2)$ and $-a(y_2 - y_1)$, respectively, as if there were a spring stretched between them. The equations then become

$$m_1 \frac{d^2 y_1}{dt^2} + (k_1 + a)y_1 - ay_2 = 0,$$
$$m_2 \frac{d^2 y_2}{dt^2} + (k_2 + a)y_2 - ay_1 = 0.$$

As a matter of convenience in the calculation, we shall introduce changes of notation: let $y_1\sqrt{m_1} = x_1$, $y_2\sqrt{m_2} = x_2$, $(k_1 + a)/m_1 = \omega_1^2$, $(k_2 + a)/m_2 = \omega_2^2$, $a/\sqrt{m_1m_2} = c$. Then the equations are

$$\begin{aligned}\frac{d^2x_1}{dt^2} + \omega_1^2x_1 - cx_2 &= 0, \\ \frac{d^2x_2}{dt^2} + \omega_2^2x_2 - cx_1 &= 0.\end{aligned}\tag{1}$$

These are two simultaneous differential equations, and there are several ways of solving them. First we may take advantage of their property of being linear with constant coefficients, and see if we cannot get exponential solutions. We assume $x_1 = Ae^{i\omega t}$, $x_2 = Be^{i\omega t}$, where A , B , ω are to be determined. Substituting, we have

$$\begin{aligned}(-\omega^2 + \omega_1^2)A - cB &= 0, \\ -cA + (-\omega^2 + \omega_2^2)B &= 0.\end{aligned}\tag{2}$$

If we regarded ω as being known, these would be two simultaneous equations for the two constants A and B . Evidently they are linear homogeneous equations. Now it is a theorem of algebra that in general two such equations do not have any solutions, unless the determinant of coefficients,

$$\begin{vmatrix} (-\omega^2 + \omega_1^2) & -c \\ -c & (-\omega^2 + \omega_2^2) \end{vmatrix}\tag{3}$$

is equal to zero. Let us see what this means. We could solve the first equation for A in terms of B : $A = Bc/(-\omega^2 + \omega_1^2)$. But we could do the same with the second, $A = B(-\omega^2 + \omega_2^2)/c$. If these solutions are to be consistent, it must be that the two factors on the right are equal, $c/(-\omega^2 + \omega_1^2) = (-\omega^2 + \omega_2^2)/c$, or $(-\omega^2 + \omega_1^2)(-\omega^2 + \omega_2^2) - c^2 = 0$. But this is just the equation obtained by setting the determinant equal to zero, so that we have verified the result of algebra. Now the equation which we have obtained, called the secular equation, can be satisfied, for we still have ω at our disposal. Solving the quadratic, this gives

$$\omega^2 = \frac{\omega_1^2 + \omega_2^2}{2} \pm \sqrt{\frac{(\omega_1^2 - \omega_2^2)^2}{4} + c^2}.\tag{4}$$

This gives two values for ω^2 , or two different possible frequencies of motion for the system. This is natural, since we should have

two frequencies if they were uncoupled, one for the one particle, the other for the other. Suppose the first, with the + sign, is called ω' , and the second, with the - sign, ω'' . It is interesting to find ω' and ω'' , in the case where c , measuring the interaction between the particles, is small. Then we can expand by the binomial theorem, obtaining

$$\begin{aligned}\omega'^2 &= \omega_1^2 + \frac{c^2}{\omega_1^2 - \omega_2^2} + \dots, \\ \omega''^2 &= \omega_2^2 + \frac{c^2}{\omega_2^2 - \omega_1^2} + \dots,\end{aligned}\quad (5)$$

showing that the frequencies approach the natural frequencies of the separate systems when the coupling goes to zero, but that they differ from them by quantities which increase as c increases. It is interesting to see that the frequencies are always spread apart by the interaction: if $\omega_1^2 > \omega_2^2$, then $\omega'^2 > \omega_1^2$, $\omega''^2 < \omega_2^2$, and correspondingly if the situation is reversed. There are several relations between ω' and ω'' which we shall need, and which we write for reference; they are easily proved from the solutions already found, and hold independently of the size of c :

$$\begin{aligned}\omega'^2 \omega''^2 &= \omega_1^2 \omega_2^2 - c^2 \\ \omega'^2 + \omega''^2 &= \omega_1^2 + \omega_2^2, \\ (-\omega'^2 + \omega_1^2)(-\omega''^2 + \omega_1^2) &= -c^2.\end{aligned}\quad (6)$$

Having determined the two possible frequencies of vibration of the system, we next find the amplitudes A' and B' corresponding to ω' , and A'' and B'' corresponding to ω'' . These are evidently given by

$$\begin{aligned}\frac{A'}{B'} &= \frac{c}{(-\omega'^2 + \omega_1^2)}, \\ \frac{A''}{B''} &= \frac{c}{(-\omega''^2 + \omega_1^2)}.\end{aligned}\quad (7)$$

That is to say, the ratios of A 's to B 's are determined, but not the values themselves. The situation is then the following: we have one possible solution, $x_1 = A'e^{i\omega't}$, $x_2 = B'e^{i\omega't}$, where the ratio of the amplitudes of x_1 and x_2 is fixed, but the magnitudes are otherwise arbitrary. Of course, there is a similar solution with $-i\omega't$ in the exponent, so that combining these in the usual way we have an arbitrary phase and amplitude, or two arbitrary constants. Next we have also the solutions $x_1 = A''e^{i\omega''t}$, $x_2 = B''e^{i\omega''t}$, of the same sort. And now, on account of the linear

nature of the equations, we can make linear combinations of these, obtaining

$$\begin{aligned}x_1 &= A'e^{i\omega't} + A''e^{i\omega''t}, \\x_2 &= B'e^{i\omega't} + B''e^{i\omega''t}.\end{aligned}\tag{8}$$

That is, each coordinate has two periods in its motion, or is doubly periodic. Since the amplitudes are to a certain extent arbitrary, it is possible for only one frequency to be excited at a time, or for both to go simultaneously.

It is interesting to consider the physical nature of the motions described by these equations. Let us assume that the two systems are only loosely coupled together (c is small). Then one possible mode of vibration has frequency ω' , only slightly greater than the frequency ω_1 which the first oscillator would have had without coupling. It is not a vibration of the first oscillator alone; both are vibrating at the same time. However, if we examine the coefficients A' , B' in this case, we find that B' is small compared with A' , meaning that the amplitude of the second oscillator is small compared with that of the first. Thus, using $B'/A' = (\omega_1^2 - \omega'^2)/c$, and $\omega'^2 = \omega_1^2 + c^2/(\omega_1^2 - \omega_2^2) + \dots$, we have approximately $B'/A' = c/(\omega_2^2 - \omega_1^2)$. This is as if the first oscillator, vibrating with frequency ω' , which is approximately ω_1 , and amplitude A' , were forcing the second oscillator by virtue of the coupling, with a force cx_1 , or $cA'e^{i\omega't}$, or approximately $cA'e^{i\omega_1 t}$. This would produce a forced amplitude of $(cA'e^{i\omega't})/(\omega_2^2 - \omega'^2)$, which is just what we have found. Similarly the second oscillator can vibrate almost by itself, with the frequency ω'' which almost equals ω_2 , but it reacts back on the first and produces a small forced amplitude. It is now in the further approximations to the interaction that the differences between ω_1 and ω' , ω_2 and ω'' , come in.

We have considered the types of vibrations separately. But there is no reason why both cannot be simultaneously excited, so that each particle will be vibrating with both periods at once. Then the phenomenon of beats can easily come in; for the sum of two sinusoidal vibrations of different frequencies is equivalent to a single vibration of varying amplitude, as we see from the equation

$$\cos \omega't + \cos \omega''t = \left(2 \cos \frac{\omega' - \omega''}{2}t \right) \cos \frac{\omega' + \omega''}{2}t,$$

where the first expression, in parentheses, represents an amplitude

oscillating with the slow frequency $(\omega' - \omega'')/2$, and modulating the latter term, a rapid vibration of frequency $(\omega' + \omega'')/2$. If ω' and ω'' are approximately equal, the effect gets most marked, the frequency of the beats approaching zero. There is in this case a pulsation of amplitude and energy from one of the oscillators to the other. This is often seen in other similar problems. Thus, if a weight is hung from a spiral spring and is set vibrating up and down, it will be observed that after a certain lapse of time the vertical motion will decrease, but there will be a torsional motion of considerable amplitude. As time goes on, these two forms of motion will alternately take up large amplitudes. The reason is that there is a coupling between the two forms of oscillation, and the beat phenomenon we have just described comes into play.

71. Normal Coordinates.—We have just seen that the general oscillation of two coupled particles is a sum of two vibrations of different frequencies. If only one of these vibrations is excited, both particles oscillate with the same frequency but different amplitudes. It now proves to be possible to introduce new coordinates X and Y , called normal coordinates, given by linear combinations of the displacements x_1 and x_2 of the two particles, which have the following properties: the generalized force acting on X is proportional to X alone, independent of Y , so that the equations of motion are separated, and X and Y execute independent simple harmonic vibrations, of different frequencies. When one of the coordinates alone is different from zero, the other remaining equal to zero, just one of the two vibrations is excited. The existence of such coordinates is made plausible by the following fact: if one vibration alone is excited, x_1 is proportional say to α times a sinusoidal function of time, x_2 to β times the same sinusoidal function. In this case $\beta x_1 - \alpha x_2$ will be always zero. This linear combination of x_1 and x_2 will be proportional, then, to the normal coordinate associated with the second type of vibration, which is not excited in the case mentioned. By assuming that the second vibration alone is excited, we can in a similar way infer the form of the first normal coordinate. We proceed in the next paragraph to the general formulation of the normal coordinates.

Suppose we set up quantities X , Y , defined by the equations

$$\begin{aligned} x_1 &= \alpha' X + \alpha'' Y, \\ x_2 &= \beta' X + \beta'' Y, \end{aligned} \tag{9}$$

where $C\alpha' = A'$, $C\beta' = B'$, $D\alpha'' = A''$, $D\beta'' = B''$, C and D being constants. Since only the ratios of the α 's and β 's are so far determined, we may demand that the magnitudes be so fixed in this case that $\alpha'^2 + \beta'^2 = 1$, $\alpha''^2 + \beta''^2 = 1$. This is called the condition of normalization, and we shall see its significance a little later. Our quantities X and Y can now be treated as generalized coordinates, and we can easily see that the equations of motion, in terms of them, have the variables separated. Let us set up the equations of motion in these new variables. We have

$$\begin{aligned} T &= \frac{1}{2} \left[\left(\frac{dx_1}{dt} \right)^2 + \left(\frac{dx_2}{dt} \right)^2 \right] \\ &= \frac{1}{2} \left[\left(\alpha' \frac{dX}{dt} + \alpha'' \frac{dY}{dt} \right)^2 + \left(\beta' \frac{dX}{dt} + \beta'' \frac{dY}{dt} \right)^2 \right] \\ &= \frac{1}{2} \left[(\alpha'^2 + \beta'^2) \left(\frac{dX}{dt} \right)^2 + (\alpha''^2 + \beta''^2) \left(\frac{dY}{dt} \right)^2 + \right. \\ &\quad \left. (\alpha' \alpha'' + \beta' \beta'') \frac{dX}{dt} \frac{dY}{dt} \right]. \end{aligned}$$

Using the relations (6) and (7), the last term can be shown to be zero. This is called the condition of orthogonality, for reasons which will later be evident. Using the normalization conditions mentioned above, we have finally

$$T = \frac{1}{2} \left[\left(\frac{dX}{dt} \right)^2 + \left(\frac{dY}{dt} \right)^2 \right]. \quad (10)$$

Next for the potential energy we have, from the original equations,

$$\begin{aligned} V &= \frac{1}{2} (\omega_1^2 x_1^2 + \omega_2^2 x_2^2 - 2c x_1 x_2), \\ &= \frac{1}{2} [(\omega_1^2 (\alpha' X + \alpha'' Y)^2 + \omega_2^2 (\beta' X + \beta'' Y)^2 \\ &\quad - 2c (\alpha' X + \alpha'' Y) (\beta' X + \beta'' Y)] \\ &= \frac{1}{2} \{ (\omega_1^2 \alpha'^2 + \omega_2^2 \beta'^2 - 2c \alpha' \beta') X^2 \\ &\quad + (\omega_1^2 \alpha''^2 + \omega_2^2 \beta''^2 - 2c \alpha'' \beta'') Y^2 \\ &\quad + 2[\omega_1^2 \alpha' \alpha'' + \omega_2^2 \beta' \beta'' - c(\alpha' \beta'' + \alpha'' \beta')] X Y \}. \end{aligned}$$

Here it can be shown by a little manipulation that the first parenthesis equals ω'^2 , the second ω''^2 , and the third is zero, so that

$$V = \frac{1}{2} (\omega'^2 X^2 + \omega''^2 Y^2). \quad (11)$$

In terms of the new variables, the variables are separated, and

Lagrange's equations become simply $d^2X/dt^2 + \omega'^2X = 0$, $d^2Y/dt^2 + \omega''^2Y = 0$, whose solutions are $X = \text{constant} \times e^{i\omega't}$, $Y = \text{constant} \times e^{i\omega''t}$. Thus each of the generalized coordinates executes a simple harmonic motion, which of course can have arbitrary amplitude and phase, and our final result, if we set the first constant equal to C , the second to D , is

$x_1 = \alpha'X + \alpha''Y = \alpha'(Ce^{i\omega't}) + \alpha''(De^{i\omega''t}) = A'e^{i\omega't} + A''e^{i\omega''t}$, etc., agreeing with the results already found.

It may be proved in general that for any mechanical problem in which the potential is a quadratic function of the coordinates, coordinates of this kind (called normal coordinates) can be set up, having the property that they have no cross terms between different coordinates in either the kinetic or the potential energy, so that the Lagrangian function is a sum of squares of coordinates and velocities, with constant coefficients, and the variables are separated in the Lagrangian equations. The general method of setting up these normal coordinates follows exactly the model we have found for our simple problem. This is one of the few sorts of mechanical problems in which a general solution is possible, for no such theorem holds with other laws of force. The equations of motion for the normal coordinates are just like those for harmonic oscillators, so that their solutions are sinusoidal vibrations. In general, there are then as many fundamental periods in the motion as there are constants, so that the motion is multiply periodic.

The normal coordinates are of particular value when we come to discuss the action of external forces on the coupled systems. For suppose there are external forces F_1 and F_2 acting on the two particles respectively, in addition to the elastic forces already considered. Then we can set up the generalized forces acting on the two normal coordinates, by the method described in Chap. VIII. If these are F_X and F_Y , we have

$$F_X = F_1 \frac{\partial x_1}{\partial X} + F_2 \frac{\partial x_2}{\partial X} = \alpha'F_1 + \beta'F_2,$$

$$F_Y = F_1 \frac{\partial x_1}{\partial Y} + F_2 \frac{\partial x_2}{\partial Y} = \alpha''F_1 + \beta''F_2.$$

Then the equations of motion are simply

$$\frac{d^2X}{dt^2} + \omega'^2X = F_X,$$

$$\frac{d^2 Y}{dt^2} + \omega'^2 Y = F_Y, \quad (12)$$

showing that these normal coordinates have the same sort of equations of motion, under the action of external forces, as single oscillators. Thus the complete solution will be a sum of a particular solution of the inhomogeneous equations, consisting of vibrations of the same nature as the external force, capable, therefore, of showing resonance phenomena, and of a general solution of the homogeneous equations, of the sort we have found.

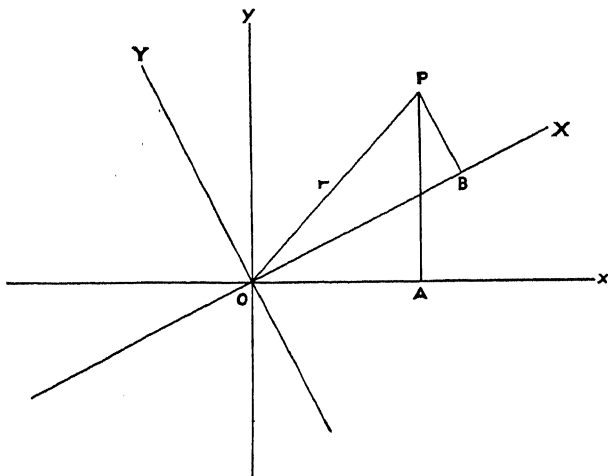


FIG. 18.—Rotation of coordinates. The distances OA and PA are the x and y coordinates of the point P , and OB and PB are the X and Y coordinates.

Under certain circumstances, a damping force proportional to the velocity will also be expressed in terms of normal coordinates as a constant times the time rate of change of the normal coordinate, but this is not always true. We shall discuss this question in the next chapter.

72. Relation of Problem of Coupled Systems to Two-dimensional Oscillator.—Our problem of two coupled one-dimensional oscillators reminds us strongly of the case of two-dimensional oscillators encountered in Chap. IX. Here, as there, we have two coordinates (x_1 and x_2 here, x and y there), and linear restoring forces. But the difference is that here the restoring force acting on each coordinate depends on the values of both. The corresponding problem in the two-dimensional case would be that where $F_x = -ax + cy$, $F_y = -by + cx$, where $a = \omega_1^2$,

$b = \omega_2^2$. And obviously the problem can be solved just as we have treated our case of the coupled oscillators. That is, we introduce new variables X, Y , defined by the equations $x = \alpha'X + \alpha''Y$, $y = \beta'X + \beta''Y$, where the α 's and β 's have the values found above, and in terms of the new variables X and Y we have separation, and get a solution in which X and Y execute periodic vibrations of different frequencies.

But now we can get a very simple geometrical interpretation of our change of variables: it is merely a rotation of coordinates. To see this, let us first consider what a rotation of coordinates means analytically. In Fig. 18, we see old coordinates xy , and new, rotated ones, XY . The xy and XY coordinates of a point P are indicated. Now there is a very simple vector way of writing the coordinates. Let i, j be the unit vectors along x and y , respectively, and I, J along X, Y . Further, let r be the radius vector from the origin to point P . Then evidently we have $x = (i \cdot r)$, $y = (j \cdot r)$, $X = (I \cdot r)$, $Y = (J \cdot r)$. But we can express i and j in terms of I and J , or vice versa:

$$\begin{aligned} i &= (i \cdot I)I + (i \cdot J)J, \\ j &= (j \cdot I)I + (j \cdot J)J. \end{aligned}$$

Hence we have

$$\begin{aligned} x &= (i \cdot r) = (i \cdot I)(I \cdot r) + (i \cdot J)(J \cdot r) \\ &= (i \cdot I)X + (i \cdot J)Y, \end{aligned} \tag{13}$$

and

$$y = (j \cdot I)X + (j \cdot J)Y.$$

These are linear equations of just the sort already found and agree if

$$\begin{aligned} (i \cdot I) &= \alpha', \\ (i \cdot J) &= \alpha'', \\ (j \cdot I) &= \beta', \\ (j \cdot J) &= \beta''. \end{aligned} \tag{14}$$

We may not assume, however, that any linear transformation of this sort corresponds to a rotation; the general transformation would be to a stretched, oblique set of axes. For the new coordinates to be obtained from the old by merely rotating, we must have two conditions: (1) the vectors I and J must be at right angles, or orthogonal, to each other; (2) I and J must be of unit length, or, as we say, normalized. That is, in vector notation,

$(I \cdot J) = 0$, $I^2 = J^2 = 1$. Now we can express these equations by taking components along the x , y axes: since $I = (i \cdot I)i + (j \cdot I)j$, $J = (i \cdot J)i + (j \cdot J)j$,

$$\begin{aligned} I \cdot J = 0 &= (i \cdot I)(i \cdot J) + (j \cdot I)(j \cdot J) \\ &= \alpha' \alpha'' + \beta' \beta'', \end{aligned} \quad (15)$$

or the orthogonality conditions, which we have already seen to be satisfied, and whose significance we now see. Also

$$\begin{aligned} I^2 = 1 &= (i \cdot I)^2 + (j \cdot I)^2 = \alpha'^2 + \beta'^2 \\ J^2 = 1 &= \alpha''^2 + \beta''^2, \end{aligned} \quad (16)$$

or the normalization conditions, which we satisfied by proper choice of arbitrary constants. We can, in conclusion, make the following statement: any linear transformation in which the transformation coefficients satisfy the orthogonality and normalization conditions corresponds to a rotation of coordinates.

The advantage of making our rotation is seen when we consider the mechanical problem. In the original problem we have force components $F_x = -ax + cy$, $F_y = -by + cx$. We can find the components of force in the new variables. Evidently

$$\begin{aligned} F_x &= (F \cdot i), F_y = (F \cdot j), \text{ and similarly} \\ F_x &= (F \cdot I) = (F \cdot i)(i \cdot I) + (F \cdot j)(j \cdot I) = \alpha' F_x + \beta' F_y \\ &= \alpha'(-ax + cy) + \beta'(-by + cx) \\ &= -(\alpha'a - \beta'c)x - (-\alpha'c + \beta'b)y \\ &= -(\alpha'a - \beta'c)(\alpha'X + \alpha'Y) - (-\alpha'c + \beta'b)(\beta'X + \beta'Y) \\ &= -(\alpha'^2a - 2\alpha'\beta'c + \beta'^2b)X - \\ &\quad (\alpha'\alpha''a - \alpha'\beta'c - \alpha'\beta''c + \beta'\beta''b)Y. \end{aligned}$$

But by results already proved, we easily see that the first parenthesis equals ω'^2 (or a corresponding expression in terms of a and b), and the second is zero, so that $F_x = -\omega'^2 X$, and similarly F_y turns out to be $-\omega''^2 Y$. In other words, by this rotation of axes, we have got each component of force to depend on displacement in that direction alone. Incidentally, the method of finding the components of a vector in rotational coordinates which we have used is of general application.

The object of the rotation becomes even clearer when we consider the potential energy. This is the quantity whose x derivative is $ax - cy$, and y derivative is $by - cx$. First we note that $\partial F_x / \partial y = \partial F_y / \partial x = c$, so that the curl of the force is zero, and the potential exists. Then we easily see that $V =$

$\frac{1}{2}(ax^2 + by^2 - 2cxy)$, or $\frac{1}{2}(\omega_1^2x^2 + \omega_2^2y^2 - 2cxy)$. An equipotential, obtained by setting this expression equal to a constant, is an ellipse with its center at the origin, but with its major and minor axes inclined at an angle to the xy axes, unless $c = 0$. But now we have seen that the potential in the new coordinates has the expression $V = \frac{1}{2}(\omega'^2X^2 + \omega''^2Y^2)$. If this is equal to a constant, the result is the equation of an ellipse whose principal axes are along the X and Y axes. In other words, our whole change of variables has been a rotation of the coordinate axes to point along the principal axes of the elliptical equipotentials. The process of rotating axes to coincide with the principal axes of an ellipse or ellipsoid is a common thing in mathematical physics. We have already seen one example in the last chapter, where we had the ellipsoid of inertia, and used the principal axes as coordinates. Other illustrations come from the theory of elasticity, where there is an ellipsoid of stress at each point, and we often use the principal axes of stress as coordinates. Again, in wave mechanics, examples of the same sort of process are constantly found.

73. The General Problem of the Motion of Several Particles.—

The present problem is the first one we have met in which there are several particles interacting with each other, and it has illustrated one of the useful methods of attack on such a problem. This is to take all the coordinates, whether they refer to one or another particle, and imagine them all plotted in a many-dimensional space, like the phase space which we discussed in connection with the Hamiltonian method, but with only enough dimensions to take care of coordinates, not of momenta. Such a space is often called a configuration space. Then the motion of the system is given by the motion of a point in configuration space. If there is a potential, it is a function of position in configuration space. We can then apply many of the same ideas to the motion of the point in many-dimensional space that we would to the motion of a single particle in three-dimensional space. Thus there will be parts of configuration space where $E - V$ is positive; there the point can go, but it cannot enter the regions where $E - V$ is negative. In some cases, changes of variables in configuration space can simplify the problem enough so that we can separate variables, or at least go far toward a solution. The present chapter has supplied one instance. Another is found in the problem of two particles, as

the earth and sun, exerting forces on each other but not being acted on by outside bodies. There we can introduce new coordinates: first, the three coordinates of the center of gravity of the system; second, the coordinates of one particle relative to the other. And in terms of these new coordinates, the three coordinates of the center of gravity become separated from the others, resulting in a uniform motion of the center of gravity in a straight line, and the relative motion reduces to a problem mathematically equivalent to the motion of a single particle in three-dimensional space. The changes of variables used in these cases generally have the property, which we have noted in the present case, of mixing up the coordinates of two or more particles in a single generalized coordinate.

Problems

1. Two balls, each of mass m , and three weightless springs, one of length $2d$, the others of length d , are connected together in the arrangement spring d —ball—spring $2d$ —ball—spring d , and the whole thing is stretched in a straight line between two points, with a given tension in the spring. Gravity is neglected. Investigate the small vibrations of the balls at right angles to the straight line, assuming motion only in one plane. Show in general that there are two modes of vibration, one having the lower frequency, in which both balls oscillate to the same side at one time, then the other, and the second mode, with higher frequency, where they oscillate to opposite sides. (Hint: if the first is displaced x_1 , and the second x_2 , and if these displacements are so small that the tension t is unchanged, then there will be two forces acting on the first ball: a force t toward the point of support, making an angle whose tangent is x_1/d , and another directed toward the second ball, at an angle whose tangent is $(x_2 - x_1)/2d$. The component at right angles to the straight line, and thus producing the motion, is then $-x_1(t/d) + (x_2 - x_1)(t/2d)$. Similarly the force on the second is $-x_2(t/d) + (x_1 - x_2)(t/2d)$.

2. Assume two resistanceless circuits, one with L_1 , C_1 , the other with L_2 , C_2 , coupled together by having a mutual inductance M between the two inductances (that is, back e.m.f. of self- and mutual inductance is $-L_1 di_1/dt - M di_2/dt$ in the first circuit, and $-L_2 di_2/dt - M di_1/dt$ in the second circuit, where i_1 , i_2 are the currents in the circuits). Find the frequencies of the natural oscillations of the coupled system.

3. In Prob. 2, assume that the circuits have small resistances R_1 and R_2 , respectively, so small that the logarithmic decrements of the separate circuits are small. Discuss the damped oscillations, showing that the solution can be carried out if squares of resistances are small enough to be neglected, but that it leads to a biquadratic equation for the frequency for large R .

4. Two identical pendulums hang from a support which is slightly yielding, so that they can interchange energy. Assume that coupling is linear. Now suppose one pendulum is set into motion, the other being at rest. Show that gradually the first pendulum will come to rest, the second taking

up the motion, and that there is a periodic pulsation of the energy from one pendulum to the other. Show that the frequency of this pulsation gets smaller as the coupling becomes smaller, until with an infinitely rigid support the energy remains always in the first pendulum (this is all without damping forces).

5. One simple pendulum is hung from another; that is, the string of the lower pendulum is tied to the bob of the upper one. Discuss the small oscillations of the resulting system, assuming arbitrary lengths and masses. Use the angles which each string makes with the vertical as generalized coordinates. In the special case of equal masses and equal lengths of strings, show that the frequencies of the motion are given by $\sqrt{g(2 + \sqrt{2})/L}$.

6. Show that if the mass of the upper pendulum becomes very great compared with the lower one, the solution of Prob. 5 approaches that of Prob. 8, Chap. IV. Show in the other limiting case, where the upper mass is small compared with the lower one, that the motion consists approximately of an oscillation of the large mass with a period derived from the combined length of both pendulums, and a more rapid oscillation of the small mass back and forth with respect to the line connecting point of support and large mass.

7. Given an ellipse $ax^2 + bxy + cy^2 = d$, perform a rotation of axes so that the new coordinates will lie along the major and minor axes of the ellipse. From this rotation, find the angle between the major axis and the x axis, in terms of the coefficients a , b , c , d . It is simplest to write the transformation directly in terms of the angle θ : $x' = x \cos \theta + y \sin \theta$, etc.

8. Show that if the equations

$$x' = a_{11}x + a_{12}y + a_{13}z,$$

$$y' = a_{21}x + a_{22}y + a_{23}z,$$

$$z' = a_{31}x + a_{32}y + a_{33}z$$

represent a rotation of coordinates, the a 's satisfy orthogonality and normalization relations, both of the form $a_{11}a_{12} + a_{21}a_{22} + a_{31}a_{32} = 0$, $a_{11}^2 + a_{21}^2 + a_{31}^2 = 1$, and of the form $a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23} = 0$, $a_{11}^2 + a_{12}^2 + a_{13}^2 = 1$.

9. In the rotation of coordinates above, show that the inverse transformation is given by

$$x = a_{11}x' + a_{21}y' + a_{31}z',$$

$$y = a_{12}x' + a_{22}y' + a_{32}z',$$

$$z = a_{13}x' + a_{23}y' + a_{33}z'.$$

Prove that the determinant of the a 's is equal to unity.

10. Find the components of an arbitrary vector in the rotated set of coordinates given in Prob. 8. Show that the components of grad V , where V is a scalar, in the rotated axes, are $\partial V/\partial x'$, $\partial V/\partial y'$, $\partial V/\partial z'$; that is, that the gradient is invariant under a rotation of axes (has the same form in the new axes as in the old).

11. Prove that the divergence, curl, and Laplacian are invariant under a rotation.

12. Set up a method for getting the direction cosines of the principal axes of inertia of a body, and the values of the principal moments of inertia, if the moments and products of inertia are known in a particular coordinate system.

CHAPTER XII

THE VIBRATING STRING, AND FOURIER SERIES

In this chapter we turn to the discussion of the motion of a continuous medium. There are examples of such motion in one, two, or three dimensions; as a vibrating string in one dimension, a membrane in two, and an elastic solid, or gas, in three dimensions. We first consider the motion of a one-dimensional body, or string. Suppose we have a string of length L , mass μ per unit length (constant), with a tension T , set into transverse vibrations. From our elementary work, we know that an infinite number of modes of vibrations, or overtones, are possible. For the n th overtone, if it is present alone, the shape of the string at any time is given by $\sin(n\pi x/L)$, where x is the coordinate of a point on the string measured from one end, and the function is proportional to the displacement transverse to the string. The frequency of this overtone is $\omega_n/2\pi$, where $\omega_n = (n\pi/L)\sqrt{T/\mu}$. Thus if A_n is the complex amplitude of this overtone, and u is the displacement of the point x , we have

$$u = \text{real part of } \sum_{n=1}^{\infty} A_n \sin \frac{n\pi x}{L} e^{i\omega_n t},$$

where we sum over all the possible overtones. Our first task is to derive these results from fundamental principles.

74. Differential Equation of the Vibrating String.—Assume that at a given time the string is displaced so that its shape is given by $u(x)$. We consider how this curve will change with time, and consider transverse displacements so small that the tension T may be considered constant throughout the string. Take a short element of the string of length dx and mass μdx . Its acceleration is $\partial^2 u / \partial t^2$ (x kept constant), so that its mass times its acceleration is $\mu dx \partial^2 u / \partial t^2$. This must be equal to the force acting on this element which arises from the tensions. These tensions (which we take equal to each other in magnitude) would cancel each other exactly if the string were straight, but when it is curved, they each give rise to components approxi-

mately perpendicular to the string which vary with the curvature of the string (see Fig. 19). At any point x , this component is approximately $T \partial u / \partial x$, and we work only to the approximation

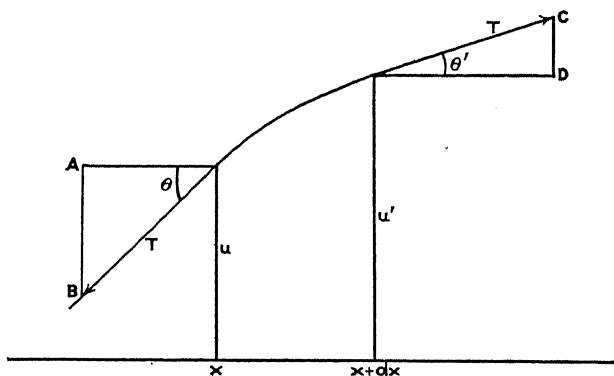


FIG. 19.—Tensions on an element of string. Vertical component at x is $-T \sin \theta$. If we approximate $\sin \theta$ by $\tan \theta$, this is $-T \frac{\partial u}{\partial x}$. Similarly at $x + dx$, the component is $+T \frac{\partial u}{\partial x}$, but now computed at $x + dx$.

to which this is true. Thus the total force on the element of string is

$$T \left[\left(\frac{\partial u}{\partial x} \right)_{x+dx} - \left(\frac{\partial u}{\partial x} \right)_x \right] = T \frac{\partial^2 u}{\partial x^2} dx,$$

if we expand the first term in a Taylor's series and retain only the first two terms of the expansion. Thus our equation of motion is

$$\mu \frac{\partial^2 u}{\partial t^2} dx = T \frac{\partial^2 u}{\partial x^2} dx,$$

or

$$\mu \frac{\partial^2 u}{\partial t^2} = T \frac{\partial^2 u}{\partial x^2}. \quad (1)$$

This is a partial differential equation, since it contains partial derivatives. This appearance of partial derivatives is characteristic of all equations of motion of continuous media. Since the equation is linear, with constant coefficients, let us try to solve it by the exponential method, assuming $u = e^{i(\omega t + kx)}$, as would be suggested by the solution in terms of overtones. The equation of motion leads immediately to $-\mu\omega^2/T = -k^2$, determining ω in terms of k . Combining two exponential solu-

tions, allowable since the equation is linear and homogeneous, we have

$$u = Ae^{i\omega t} \sin kx,$$

or

$$u = Be^{i\omega t} \cos kx. \quad (2)$$

Now we must introduce the boundary conditions, which tell us that the string is held fixed at both ends, so that $u = 0$ when $x = 0$, and when $x = L$. From the first of these conditions $B = 0$, and we take only the sine function. From the second, we must have $\sin kL = 0$, or $kL = n\pi$,

$$k = \frac{n\pi}{L},$$

where $n = 1, 2, 3, \dots$

Hence the solution is

$$u = A_n e^{i\omega_n t} \sin \frac{n\pi x}{L}, \quad (3)$$

where

$$\omega_n = \frac{n\pi}{L} \sqrt{\frac{T}{\mu}}.$$

Superposing the solutions of all the different n 's, as we may from the nature of the differential equation, we obtain the solution mentioned at the beginning of the chapter.

Our differential equation is a linear homogeneous partial differential equation of second order. As such, any linear combination of solutions is itself a solution. But now we have, not a small number of arbitrary constants, but the doubly infinite set A_n , $n = 1, 2, \dots$ (A_n is complex). This is characteristic of all partial differential equations. Sometimes instead of having an infinite set of arbitrary constants, we have an arbitrary function. In our case the A 's are determined by giving the amplitude and phase of each overtone. These must be determined from the initial conditions; that is, from the values of $u(x)$ and $\dot{u}(x)$ at $t = 0$. The essential point is that our partial differential equation is equivalent to an infinite number of ordinary differential equations, so that we need an infinite number of constants.

75. The Initial Conditions for the String.—Suppose we wish to satisfy initial conditions of the following sort for a vibrating string: at $t = 0$, the displacement and velocity are given functions of x . That is, if the displacement is $u(x, t)$, then $u(x, 0) =$

$f(x)$, $\frac{\partial u}{\partial t}(x, 0) = F(x)$, where $f(x)$, $F(x)$, are arbitrary functions.

Now we may write

$$u(x, t) = \sum_n (C_n \cos \omega_n t + D_n \sin \omega_n t) \sin \frac{n\pi x}{L}, \quad (4)$$

using the real form of the function of time, and

$$u(x, t) = \sum_n (-C_n \omega_n \sin \omega_n t + D_n \omega_n \cos \omega_n t) \sin \frac{n\pi x}{L}.$$

Thus we must have

$$\begin{aligned} u(x, 0) = f(x) &= \sum_n C_n \sin \frac{n\pi x}{L}; \\ u(x, 0) = F(x) &= \sum_n D_n \omega_n \sin \frac{n\pi x}{L}. \end{aligned} \quad (5)$$

To satisfy either of these conditions, we must be able to expand our arbitrary function in series of sines, and to find the coefficients C_n or D_n of these expansions. Having found the coefficients, we can at once set up the series for $u(x, t)$. This is a special case of Fourier expansion, and we now proceed to consider the general problem of Fourier series, a question of general interest apart from the application to a string.

76. Fourier Series.—We shall state Fourier's theorem. Given an arbitrary function $\phi(x)$. Then [unless $\phi(x)$ contains an infinite number of discontinuities in a finite range, or similarly misbehaves itself], we can write

$$\phi(x) = \frac{A_0}{2} + \sum_{n=1}^{\infty} \left(A_n \cos \frac{2n\pi x}{X} + B_n \sin \frac{2n\pi x}{X} \right),$$

where

$$A_n = \frac{2}{X} \int_{-X/2}^{X/2} \phi(x) \cos \frac{2n\pi x}{X} dx, B_n = \frac{2}{X} \int_{-X/2}^{X/2} \phi(x) \sin \frac{2n\pi x}{X} dx. \quad (6)$$

This equation holds for values of x between $-X/2$ and $X/2$, but not in general outside this range. The series of sines and cosines is called Fourier's series. Obviously a special case of it could be used in our problem of the string, the case where only the coefficients of the sine terms were different from zero.

There are two sides to the proof of Fourier's theorem. First, we may prove that, if a series of sines and cosines of this sort can represent the function, then it must have the coefficients we have given. That is simple, and we shall carry it through. But, second, we could show that the series we so set up actually represents the function. That is, we should investigate the convergence of the series, show that it does converge and that its sum is the function $\phi(x)$. This second part we shall omit, merely stating the results of the discussion.

77. Coefficients of Fourier Series.—Let us suppose that $\phi(x)$ is given by the series above, and ask what values of A 's and B 's we must have if the equation is to be true. Multiply both sides of the equation by $\cos(2m\pi x/X)$, where m is an integer, and integrate from $-X/2$ to $X/2$. We have then

$$\int_{-X/2}^{X/2} \phi(x) \cos \frac{2m\pi x}{X} dx = \int_{-X/2}^{X/2} \left\{ \frac{A_0}{2} \cos \frac{2m\pi x}{X} + \sum_n \left(A_n \cos \frac{2n\pi x}{X} \cos \frac{2m\pi x}{X} + B_n \sin \frac{2n\pi x}{X} \cos \frac{2m\pi x}{X} \right) \right\} dx.$$

But now we shall show in the next paragraph that

$$\int_{-X/2}^{X/2} \cos \frac{2n\pi x}{X} \cos \frac{2m\pi x}{X} dx = 0, \quad (7)$$

if n and m are integers, unless $n = m$, and that

$$\int_{-X/2}^{X/2} \sin \frac{2n\pi x}{X} \cos \frac{2m\pi x}{X} dx = 0,$$

if n and m are integers. Thus all terms on the right are zero but one, for which $n = m$. The first term falls in with the rule, when we remember that $\cos 0 = 1$. This one term then gives us

$$A_m \int_{-X/2}^{X/2} \cos^2 \frac{2m\pi x}{X} dx = A_m \frac{X}{2},$$

as we can readily show. Hence

$$A_n = \frac{2}{X} \int_{-X/2}^{X/2} \phi(x) \cos \frac{2n\pi x}{X} dx.$$

In a similar way, multiplying by $\sin(2m\pi x/X)$, we can prove the formula for B_n .

In our derivation of coefficients, we have used the following results: $\int_{-X/2}^{X/2} \cos \frac{2n\pi x}{X} \cos \frac{2m\pi x}{X} dx = 0$, if n, m are different integers, and similar relations with sines. We can prove these very easily from trigonometry. Thus

$$\cos a \cos b = \frac{[\cos (a + b) + \cos (a - b)]}{2},$$

so that our quantity is the integral of this, or

$$\frac{X}{2} \left[\frac{\sin \frac{2\pi(n+m)x}{X}}{2\pi(n+m)} + \frac{\sin \frac{2\pi(n-m)x}{X}}{2\pi(n-m)} \right] \Big|_{-X/2}^{X/2}.$$

But the quantity in brackets is zero at both limits, if n, m are integers, and the result is zero. Such proofs hold in the other cases. The exception, of course, is the case $n = m$, in which the integrand is $\frac{1}{2}(\cos(4\pi nx/X) + 1)$, so that, while the first term gives no contribution to the result, the second gives $\frac{1}{2} \int_{-X/2}^{X/2} dx = \frac{X}{2}$.

78. Convergence of Fourier Series.—In this section we shall merely quote results. In the first place, the series cannot in general represent the function, except in the region between $-X/2$ and $X/2$. For the series is periodic, repeating itself in every half period, while the function in general is not. Only periodic functions of this period can be represented in all their range by Fourier series. If we try to represent a nonperiodic function, the representation will be correct within the range from $-X/2$ to $X/2$, but the same thing will automatically repeat outside the range. Incidentally, we can easily change the range in which the function is correct. If we merely change the range of integration so as to be from x_0 to $x_0 + X$, where x_0 is arbitrary, the series will represent the function within this range. The case we have used above corresponds to $x_0 = -X/2$; another choice frequently made is $x_0 = 0$. Then again, if we change the value of X , we can change the length of the range in which the series is correct. To represent a function through a large range of x , we may use a large value of X .

Although the range within which a Fourier series converges to the value of the function it is supposed to represent is limited, as we have seen, there is a compensation, in that within this range

a Fourier series can be used to represent much worse curves than a power series. Thus the convergence of the series is not impaired if the function has a finite number of discontinuities. It can consist, for example, of one function in one part of the region, another in another (in this case, to carry out the integrations, we must break up the integral into separate integrals over these parts, and add them). The less serious the discontinuities, however, the better the convergence. Thus if the function itself has discontinuities, the coefficients will go off as $1/n$, while if only the first derivative has discontinuities the coefficients go off as $1/n^2$, and so on. Differentiating a function makes the convergence of a series worse, as we can see, for example, if a function is continuous but its first derivative discontinuous. Then the coefficients go off as $1/n^2$, but if we differentiate, the coefficients of the resulting series will go off as $1/n$. There is an interesting point connected with the series for a discontinuous function. If the function jumps from one value u_1 to another u_2 at a given value of x , then the series at this point converges to the mean value, $(u_1 + u_2)/2$.

79. Sine and Cosine Series, with Application to the String.—

In the special problem of the vibrating string, the series we require is somewhat different from the general case, in that there are only sines, and not cosines. We are therefore led to investigate series of sines only, or of cosines only. Suppose we take

the series $\frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos \frac{2n\pi x}{X}$, the series formed by taking

the cosine part of the general Fourier series. Now each one of the terms is even in x ; that is, if we interchange x with $-x$, the function is not changed. A cosine series represents there-

fore an even function. Similarly the sine series $\sum_{n=1}^{\infty} B_n \sin \frac{2n\pi x}{X}$,

of which each term is odd, represents an odd function (one for which, if x is interchanged with $-x$, the function changes its sign but not its magnitude). It is well known that any function $\phi(x)$ can be written as the sum of an even and an odd function: $\phi(x) = \frac{1}{2}[\phi(x) + \phi(-x)] + \frac{1}{2}[\phi(x) - \phi(-x)]$, of which the first term is even, the second odd. Thus the cosine part of a Fourier series represents the even part of the function, the sine series the odd part. As a corollary, any even function can be repre-

sented by a cosine series alone, an odd function by a sine series.

Now suppose we are really interested in a function only between 0 and $X/2$, and that we do not care what the series does outside that region. Then we may define an even function $\phi_e(x)$ as follows: it equals the given function $\phi(x)$ between 0 and

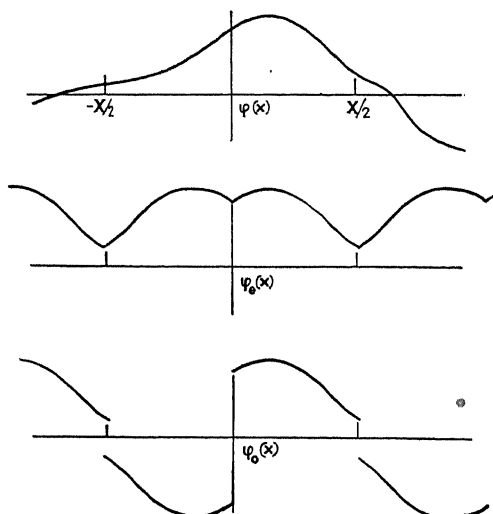


FIG. 20.—A function, with even and odd periodic functions made from it. The even and odd functions, $\phi_e(x)$ and $\phi_o(x)$, agree with the original function $\phi(x)$ between 0 and $X/2$. Between 0 and $-X/2$, $\phi_e(x)$ is the mirror image of $\phi(x)$, while $\phi_o(x)$ has the opposite sign. Outside the region from $-X/2$ to $X/2$, both functions repeat periodically with period X .

$X/2$, but has just the same value for $-x$ that it has for x (see Fig. 20). Outside the range from $-X/2$ to $X/2$, it repeats itself. The Fourier representation of ϕ_e will be a cosine series, but will represent our given function ϕ correctly between 0 and $X/2$.

Evidently it is the series $\frac{A_0}{2} + \sum_{n=1}^{\infty} A_n \cos \frac{2n\pi x}{X}$, where we write

the coefficients as the sum of two integrals,

$$\begin{aligned} A_n &= \frac{2}{X} \int_{-X/2}^{X/2} \phi_e(x) \cos \frac{2n\pi x}{X} dx \\ &= \frac{2}{X} \left(\int_{-X/2}^0 \phi(-x) \cos \frac{2n\pi x}{X} dx + \int_0^{X/2} \phi(x) \cos \frac{2n\pi x}{X} dx \right) \\ &= \frac{4}{X} \int_0^{X/2} \phi(x) \cos \frac{2n\pi x}{X} dx. \end{aligned}$$

Similarly we may define an odd function $\phi_o(x)$, which equals $\phi(x)$ between 0 and $X/2$, but at $-x$ has the negative of its value at $+x$.

This function is represented by a sine series $\sum_{n=1}^{\infty} B_n \sin \frac{2n\pi x}{X}$,

where we readily see that

$$B_n = \frac{4}{X} \int_0^{X/2} \phi(x) \sin \frac{2n\pi x}{X} dx.$$

Hence, between 0 and $X/2$, the same function can be represented by either a cosine or a sine series. But outside this range, the series represent quite different functions.

Our sine series can now be applied to the string problem. We are interested in the string between 0 and L . Let us then set $L = X/2$. The expression then becomes

$$\phi(x) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L}, \quad (8)$$

where

$$B_n = \frac{2}{L} \int_0^L \phi(x) \sin \frac{n\pi x}{L} dx.$$

This can be used first to find the coefficients C_n , from Eq. (5), substituting $u(x, 0)$ for $\phi(x)$, and next to find the quantities $D_n\omega_n$, substituting $\dot{u}(x, 0)$ for $\phi(x)$, $D_n\omega_n$ for B_n , and obtaining D_n by dividing through by ω_n . These formulas then suffice to find the constants C_n and D_n of the motion of the string, knowing the initial displacement and velocity of every point of it.

80. The String as a Limiting Problem of Vibration of Particles.

An excellent insight into the problem of the vibration of a string is obtained by regarding it as a limiting case of mechanical systems with a finite number of particles, having therefore a finite set of arbitrary constants in the solution. This is the method followed by Lagrange. Suppose we have N equal masses m at the points $x = d/2, 3d/2, \dots (N - \frac{1}{2})d$, separated by massless springs, the whole being stretched with a tension T between supports at $x = 0$ and $x = Nd = L$. This forms an approximation to the continuous string, if $\mu = m/d$, the mass per unit length. We again investigate the transverse vibrations, letting the displacement of the i th particle be u_i . The problem

is similar to Problem 1, Chap. XI. The force on the i th particle is

$$F_i = -(u_i - u_{i-1})\frac{T}{d} - (u_i - u_{i+1})\frac{T}{d},$$

except for the first particle, where we have

$$F_1 = -u_1\frac{2T}{d} - (u_1 - u_2)\frac{T}{d}$$

and for the N th,

$$F_N = -(u_N - u_{N-1})\frac{T}{d} - u_N\frac{2T}{d}.$$

Then, assuming a solution $u_i = C_i e^{i\omega t}$, we have the N equations of motion in the form

$$\begin{aligned} & \left(-m\omega^2 + \frac{3T}{d}\right)C_1 - \frac{T}{d}C_2 = 0 \\ -\frac{T}{d}C_1 & + \left(-m\omega^2 + \frac{2T}{d}\right)C_2 - \frac{T}{d}C_3 = 0 \\ -\frac{T}{d}C_2 & + \left(-m\omega^2 + \frac{2T}{d}\right)C_3 - \frac{T}{d}C_4 = 0 \\ & \dots \dots \dots \\ -\frac{T}{d}C_{N-1} & + \left(-m\omega^2 + \frac{3T}{d}\right)C_N = 0. \end{aligned} \quad (9)$$

Such a set of equations, all alike in form (except here for the first and last), are called difference equations. As in the last chapter, these have a solution only for certain values of ω , given by setting the determinant of coefficients equal to zero. The determinant is now too complicated to handle simply, wherefore we adopt another method of procedure. Suppose we let $C_j = e^{ikj}$, where k is to be determined from the equations above. All the equations except the first and last take the form

$$-\frac{T}{d}e^{ik(j-1)} + \left(-m\omega^2 + 2\frac{T}{d}\right)e^{ikj} - \frac{T}{d}e^{ik(j+1)} = 0$$

or

$$-2\frac{T}{d}\cos k + \left(-m\omega^2 + 2\frac{T}{d}\right) = 0,$$

whence

$$m\omega^2 = 2\frac{T}{d}(1 - \cos k). \quad (10)$$

That is, for any ω , we can choose a value of k by this relation, so that all the equations except the first and last are satisfied. These fall into line as well if we set up $C_0 = -C_1$, and $C_{N+1} = -C_N$, so that if these conditions are satisfied we have e^{ikj} , or equally well e^{-ikj} , or $\sin kj$ or $\cos kj$, as solutions of the equations for C_j . These conditions on C_0 and C_{N+1} are essentially boundary conditions, one at each end of the string, and we readily see that they are satisfied if we make our function zero for $x = 0$, $x = L$, as we do if it is $\sin \frac{n\pi x}{L}$, where n is an integer. That is, since x is $(j - \frac{1}{2})d$ for the j th particle, we have

$$C_j = \sin \frac{n\pi}{N} \left(j - \frac{1}{2} \right), \quad (11)$$

so that $k = n\pi/N$. We see from this form of C_j that $C_0 = -C_1$, and that only those values of n up to N give us different sets of C 's. If n is greater than N , then for each integral j we get just the same value of C_j that we had for a certain n less than N , so that the whole scheme repeats itself over and over as n increases, and we really have only N distinct solutions. Similarly in the expression for the frequency, the term $1 - \cos k = 1 - \cos (n\pi/N)$ is periodic, so that as soon as n becomes greater than N we repeat the frequencies already found. There are, then, just N solutions, each with its frequency and its complex amplitude for each particle. This fits in with the single frequency for one particle and the two which we have found for two coupled particles. For each of the N particles there is an arbitrary amplitude and phase, or arbitrary complex amplitude, so that there are just $2N$ arbitrary constants. The whole solution is the sum, as n goes from 1 to N . of the real parts of $A_n e^{i\omega_n t} \sin \frac{n\pi x}{L}$, or

$$u = \sum_{n=1}^N B_n \sin \frac{n\pi x}{L} \cos (\omega_n t - \epsilon_n). \quad (12)$$

Each one of these terms represents the amplitudes of all the particles when vibrating with a particular mode of motion, analogous to an overtone of the string. To get the amplitude of the j th particle, we set $x = (j - \frac{1}{2})d$. The angular velocity ω_n of the n th overtone is given by

$$m\omega_n^2 = 2\frac{T}{d}\left(1 - \cos \frac{n\pi}{N}\right). \quad (13)$$

81. Lagrange's Equations for the Weighted String.—The equations of motion which we have discussed above may also be obtained readily from Lagrange's method, and we shall set up expressions for the kinetic and potential energies. For the kinetic energy T_1 we have simply

$$T_1 = \frac{m}{2}(\dot{u}_1^2 + \dot{u}_2^2 + \cdots + \dot{u}_N^2),$$

and for the potential energy

$$V = \frac{T}{2d}[2u_1^2 + (u_2 - u_1)^2 + (u_3 - u_2)^2 + \cdots + 2u_N^2],$$

and the Lagrangian equations

$$\frac{d}{dt}\left[\frac{\partial(T_1 - V)}{\partial\dot{u}_i}\right] - \frac{\partial(T_1 - V)}{\partial u_i} = 0$$

lead to the equations already used.

82. Continuous String as Limiting Case.—The solution we have found for the set of particles differs in two ways from the solution for the continuous string. First, there is only a finite set of overtones, and secondly, the frequencies are determined by different formulas. Both these differences disappear when the number of particles in the fixed length L becomes infinite. To determine the limiting form of the expressions for the frequency, we develop $\cos \frac{n\pi}{N}$ in a power series for large N . We thus obtain

$$\cos \frac{n\pi}{N} = 1 - \frac{1}{2}\left(\frac{n\pi}{N}\right)^2 + \cdots,$$

so that ω_n becomes

$$\omega_n = \sqrt{\left(\frac{n\pi}{N}\right)^2 \frac{T}{md}} = \frac{n\pi}{L} \sqrt{\frac{T}{\mu}},$$

using $Nd = L$ and $m = \mu d$. This agrees with our former result. In this limiting case of infinite N (and infinitesimal d) the expressions for the kinetic and potential energies become

$$T_1 = \frac{\mu}{2} \int_0^L \dot{u}^2 dx, \text{ and } V = \frac{T}{2} \int_0^L \left(\frac{\partial u}{\partial x}\right)^2 dx, \quad (14)$$

which may also be derived directly for the case of a continuous string.

Problems

1. Taking the case of four particles on a string, derive their displacements in the four possible normal vibrations, and compute their frequencies. Compare these frequencies with the first four frequencies of the corresponding continuous string. Put in $n = N + 1$, and show how the solution reduces to one already found.

2. An actual string is composed of atoms, rather than being continuous, so that it has only a finite number of possible overtones. Assume that it consists of a single string of atoms, spaced 10^{-8} cm. apart. Let the string be 1 m. long, and at such tension that its fundamental is 100 cycles per second. Find the frequency of the highest possible harmonic, and show that it is in the infra-red region of the spectrum. Show that in this highest harmonic, successive atoms vibrate in opposite phases. Substances actually have such natural frequencies in the infra-red, and they are important in connection with their specific heat.

3. Prove that $u = \sin \omega[t - (x/v)]$ is a solution of the partial differential equation for the vibrating string, if v is chosen properly, although it does not satisfy the boundary condition that the string be held at the ends. Consider the physical meaning of this solution, and show that it represents a wave traveling down the string with velocity v .

4. Superpose the wave of Prob. 3, traveling along the $+x$ axis, and a similar one traveling in the opposite direction, and show that the sum represents a standing wave of the type discussed in this chapter.

5. Find the wave length of the waves in the string, in the solution we have found in this chapter, and verify the relation $v = n\lambda$ between wave length λ , frequency n , and velocity v .

6. Proceeding as in Prob. 5, find the velocity of a wave along the weighted string, showing that it varies with frequency. Find a formula for the variation.

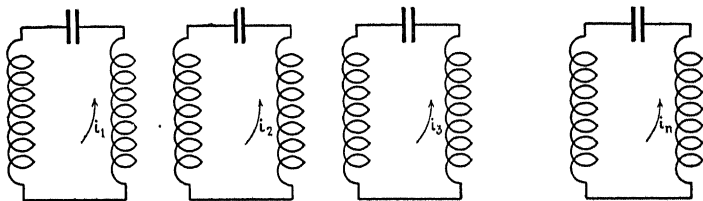


FIG. 21.—Artificial electric line.

7. An artificial electric line can be constructed according to Fig. 21, consisting of N identical resistanceless circuits, each containing inductance L , capacitance C , and coupled to each of its neighbors with mutual inductance M . Set up the differential equations for the currents i in the various circuits, showing that they reduce to the same form as with the weighted string.

8. Neglecting boundary conditions at the two ends of the line in Prob. 7, show that a disturbance can be propagated along the line with a definite velocity, as in Prob. 6.

9. A string of length L is pulled aside at a point a distance D from the end, and then released. Thus its initial shape is given by a curve made of two straight lines, and its initial velocity is zero. Find the solution for its motion, and find the amplitude of the n th harmonic.

10. Taking the solution of Prob. 9, for the special case where $D = L/2$, compute the first five terms of the Fourier series, when $t = 0$. Add them and plot the sum, showing how good an approximation they make to the correct curve.

11. A string initially at rest is struck at a distance D from the end, at $t = 0$. Find the intensity in each overtone. Approximate the initial conditions as follows: the initial displacement is zero, and the initial velocity is a constant in a small region of length d about the point D , zero elsewhere.

CHAPTER XIII

NORMAL COORDINATES AND THE VIBRATING STRING

In the preceding chapter, we have worked out the elementary theory of the vibrating string, finding the nature of the possible vibrations and the method of getting the amplitude of the overtones in terms of the initial conditions. When we begin to ask about slightly more complicated problems, however, we find that it is necessary to go further into the theory. For example, we might be interested in the nature of the forced vibrations under the action of an external sinusoidal force, or the effect of damping on the oscillations. Such questions are easily answered by introducing normal coordinates, much as we did with the two coupled oscillators. These are generalized coordinates, which prove to be closely connected with the various overtones, so that if just one normal coordinate is vibrating, that means that the string is vibrating with the corresponding pure overtone. When we write Lagrange's equations in terms of the normal coordinates, we find that we can introduce external forces easily, and solve such problems.

At the same time, the general theory of normal coordinates for vibrating strings, which we shall get into, has particularly interesting relations to many other branches of mathematical physics. We shall gain much more insight into Fourier expansion, finding a general theory of expansion of which this is a special case, but which, as we shall find later, includes expansions in Bessel's functions, spherical harmonics, and many other sorts of functions. Such problems are met not only in vibrations, but also in heat flow (for which Fourier series were originally developed), potential theory, hydrodynamics, and in the newest branch of mathematical physics, wave mechanics, or the quantum theory, used in studying atomic structure.

83. Normal Coordinates.—In Chap. XI, we investigated the vibrations of two coupled particles and set up normal coordinates to describe the motion. Since we must make a considerable extension of the idea of normal coordinates in the present chapter, it will be best to review the results we have already found. We

started with two coordinates, x_1 and x_2 , describing the displacement of the two particles. The normal coordinates X and Y were introduced by a linear transformation

$$\begin{aligned}x_1 &= \alpha'X + \alpha''Y \\x_2 &= \beta'X + \beta''Y,\end{aligned}$$

which proved to be merely a rotation of axes in the $x_1 - x_2$ space, so that X and Y were new orthogonal axes in that space. To express the fact that the transformation was just a rotation, we had certain conditions holding between the coefficients: orthogonality conditions, as $\alpha'\alpha'' + \beta'\beta'' = 0$, and normalization conditions, as $\alpha'^2 + \beta'^2 = 1$. We saw that the quantities α' , α'' , β' , β'' , had a geometrical meaning: α' was $(i \cdot I)$, and similar relations for the other quantities, showing that α' , β' , were the components along the x_1 and x_2 axes, respectively, of unit vector along X , and α'' , β'' similarly were components of unit vector along Y . The object of the rotation to normal coordinates was to separate the variables of the equation of motion, so that each normal coordinate executed a vibration of its own, as $X = Ce^{i\omega' t}$, $Y = De^{i\omega'' t}$. This was equivalent to rotation so that the new axes in the x_1x_2 space lay along the principal axes of the elliptical equipotentials of the problem.

We can now follow exactly the same model in our problem of the string. We start with the case of n weights separated by springs. By analogy, the displacement of the first weight should be a linear combination of normal coordinates, the coefficients (corresponding to the α 's and β 's) being the displacements of the first weight when only one overtone is excited, or $\sin \frac{n\pi x_1}{L}$ for the n th overtone. The displacements of these weights are taken to be $u_1 \dots u_N$. Then we set up N normal coordinates, $\phi_1 \dots \phi_N$, by the equations

$$\begin{aligned}u_1 &= \sum_{n=1}^N a_n \sin \frac{n\pi x_1}{L} \phi_n, \\u_2 &= \sum_{n=1}^N a_n \sin \frac{n\pi x_2}{L} \phi_n, \text{ etc.},\end{aligned}\tag{1}$$

where $x_j = (j - \frac{1}{2})d$, and the numbers a_n are determined by a condition soon to be described. Here the quantities $a_n \sin n\pi x_j/L$ correspond to the α 's and β 's of the preceding chapter.

But not only that: the coefficients still satisfy orthogonality and normalization conditions. The orthogonality conditions will be of the form

$$a_n a_m \left(\sin \frac{n\pi x_1}{L} \sin \frac{m\pi x_1}{L} + \sin \frac{n\pi x_2}{L} \sin \frac{m\pi x_2}{L} + \dots + \sin \frac{n\pi x_N}{L} \sin \frac{m\pi x_N}{L} \right) = 0, \quad (2)$$

where n, m are any two indices. This is true, as can be shown by trigonometrical manipulation, though we shall not stop to do it. Similarly the normalization will be

$$a_n^2 \left(\sin^2 \frac{n\pi x_1}{L} + \sin^2 \frac{n\pi x_2}{L} + \dots + \sin^2 \frac{n\pi x_N}{L} \right) = 1. \quad (3)$$

We can satisfy this by proper choice of a_n , since the parenthesis is a definitely determined, positive quantity. This is then the condition, called the normalization condition, for determining the constants a_n . Since we have as before an orthogonal transformation, we can again get a geometrical interpretation. We imagine an N -dimensional space, in which the quantities $u_1 \dots u_N$ are plotted as coordinates. Now our transformation of axes is equivalent to a rotation of coordinates in this N -dimensional space. The normal coordinates $\phi_1 \dots \phi_N$ represent new orthogonal axes in the space, in the sense that if $\phi_1 = 1$, all the other ϕ 's are zero, the corresponding point is displaced from the origin unit distance along the ϕ_1 axis. The quantities like $a_n \sin \frac{n\pi x_j}{L}$ represent the components in the direction of the old axes of unit vectors along the new axes. Thus the one written is the cosine of the angle between the ϕ_n and the x_j axes. The equations of motion are separated in the new coordinates, the solutions being $\phi_n = \text{constant} \times e^{i\omega_n t}$. Finally, the equipotentials, which are ellipsoids in the N -dimensional space, have principal axes in the directions which we have chosen for the normal coordinates. Thus the analogy with the two-dimensional problem is complete. The statements we have made without proof here are not very difficult to demonstrate, and some of them are taken up in problems.

We can now go one step farther, to the continuous string. Here the displacement of a point of the string is given by $u(x)$, where x measures the coordinate of the point, corresponding to the u_i for the problem of discrete weights. We introduce normal

coordinates $\phi_1, \dots, \phi_n, \dots$, (an infinite set, as there are an infinite number of points on the string), by the equation

$$u(x) = \sum_{n=1}^{\infty} a_n \sin \frac{n\pi x}{L} \phi_n. \quad (4)$$

The orthogonality conditions for the coefficients $a_n \sin \frac{n\pi x}{L}$ must now be written in terms of integrals, rather than sums; for we have terms for each value of x , from 0 to L , differing by infinitesimal amounts. Thus these conditions are

$$\int_0^L a_n a_m \sin \frac{n\pi x}{L} \sin \frac{m\pi x}{L} dx = 0, \quad (5)$$

where n and m are different integers. This can be immediately proved by evaluating the integral. Similarly the normalization condition is

$$\int_0^L a_n^2 \sin^2 \frac{n\pi x}{L} dx = 1. \quad (6)$$

which as before serves to determine a_n .

84. Normal Coordinates and Function Space.—We must now imagine a space, not of N dimensions, but of an infinite number. We cannot get an idea of what the coordinates mean, except by passing to the limit from the case of a finite number of mass points. With N points, and N dimensions, the first coordinate measures the displacement of the first mass point, and so on. Thus a point in the N -dimensional space determines all the coordinates, or in other words gives the displacements of all the masses. Now as N gets larger and larger, and we have more and more dimensions, it still remains true that a particular coordinate measures the displacement at a particular part of the string. We see that this interpretation persists to the limit of infinitely many variables: each coordinate is connected with a point of the string, and its value gives the displacement at that point. But there is now an interesting side light on the situation. A point in our infinitely many-dimensional space gives complete information about the displacement of each point of the string. That is, it gives $u(x)$, a function of x . Each point of this space is connected with a particular function, and each possible function is represented by a point of the space (of course, many points of the space refer to discontinuous functions and, therefore, are

not suitable for describing a string). On account of this property, our space is often called a function space.

The normal coordinates now represent a set of rectangular axes in function space, rotated with respect to the original coordinates. Each normal coordinate refers to a particular mode of vibration, or overtone. If just one of the normal coordinates is excited, say if $\phi_n = 1$, all the other ϕ 's being zero, the situation is represented by a certain point in function space; that is, by a certain function, giving the shape of the string. We can take the radius vector out to the point $\phi_n = 1$, all other ϕ 's = 0, and project it on one of our original coordinates. Thus the projection on the coordinate connected with the point x is $a_n \sin(n\pi x/L)$, showing that that is the displacement of this particular point of the string when this overtone alone is excited with unit amplitude. The expression $a_n \sin(n\pi x/L)$ is now a function; it is the function represented by a unit vector along the ϕ_n axis, in function space. Since the ϕ axes are orthogonal, we see that the scalar product of two such vectors along different axes is zero:

$$\int_0^L a_n a_m \sin \frac{n\pi x}{L} \sin \frac{m\pi x}{L} dx = 0,$$

where by analogy the scalar product takes this form, so that we have the orthogonality conditions and their geometrical meaning. Similarly the square of the unit vector, which is unity, is

$$\int_0^L a_n^2 \sin^2 \frac{n\pi x}{L} dx = 1,$$

the normalization condition. This immediately gives $a_n^2 L/2 = 1$, $a_n = \sqrt{2/L}$.

Now as before, when we introduce the normal coordinates, we have rotated axes in function space to make the new coordinates lie along the principal axes of the ellipsoidal equipotentials. And the equations of motion are separated, each normal coordinate vibrating with simple harmonic motion: $\phi_n = A_n e^{i\omega_n t}$. Finally, then, the motion is represented by

$$\begin{aligned} u(x) &= \sum_{n=1}^{\infty} \sqrt{\frac{2}{L}} \sin \frac{n\pi x}{L} \phi_n \\ &= \sum_{n=1}^{\infty} \sqrt{\frac{2}{L}} A_n e^{i\omega_n t} \sin \frac{n\pi x}{L}, \end{aligned}$$

agreeing with the value found previously. In Section 86, we carry out the demonstration that the equations of motion are separated in the normal coordinates, and then we apply them to the discussion of forced motion.

85. Fourier Analysis in Function Space.—When we come to the question of satisfying initial conditions, and of Fourier's series, we meet immediately close connections with function space. Fourier's theorem, stated for sine series, can be put in the following form, by introducing terms $\sqrt{2/L}$:

$$f(x) = \sum_{n=1}^{\infty} (\sqrt{L/2} B_n) \sqrt{2/L} \sin \frac{n\pi x}{L},$$

where

$$\sqrt{L/2} B_n = \int_0^L f(x) \sqrt{2/L} \sin \frac{n\pi x}{L} dx. \quad (7)$$

Now the functions $\sqrt{2/L} \sin (n\pi x/L)$ are the unit vectors in function space along the directions representing the overtones of the problem—functions which are often called the normal functions, or characteristic functions, of the problem. Thus Eq. (7) is just like a vector equation, stating that a vector ($f(x)$) is the sum of unit vectors $[\sqrt{2/L} \sin (n\pi x/L)]$ each multiplied by the component of the vector along the corresponding axis ($\sqrt{L/2} B_n$ is the component of $f(x)$ along the n th axis). To find these components, we need only project the vector $f(x)$ on the corresponding unit vector, which means taking the scalar product. But the scalar product, as we have seen, is an integral,

$$\left[f(x) \cdot \sqrt{2/L} \sin \frac{n\pi x}{L} \right] = \sqrt{L/2} B_n = \int_0^L f(x) \sqrt{2/L} \sin \frac{n\pi x}{L} dx. \quad (8)$$

Thus the formulas of Fourier's method have the simplest possible vector interpretation in function space. But we can also see that, if we had some other set of normalized orthogonal functions, we could proceed with an expansion in an analogous way. It is worth noting that by using Fourier's method, we can solve for

the normal coordinates in terms of $u(x)$: since $u(x) = \sum_{n=1}^{\infty} \sqrt{2/L}$

$\sin \frac{n\pi x}{L} \phi_n$, it is obvious that $\phi_n = \int_0^L u(x) \sqrt{2/L} \sin \frac{n\pi x}{L} dx$, the component of $u(x)$ along the n th axis.

86. Equations of Motion in Normal Coordinates.—To find the equations of motion, we must set up the Lagrangian function. Let us first write for the velocity of the string

$$\dot{u} = \dot{\phi}_1 \sqrt{2/L} \sin \frac{\pi x}{L} + \dot{\phi}_2 \sqrt{2/L} \sin \frac{2\pi x}{L} + \dots + \dot{\phi}_n \sqrt{2/L} \sin \frac{n\pi x}{L} \dots$$

and proceed to the expressions for the potential and kinetic energies. We have

$$\begin{aligned} T_1 &= \frac{\mu}{2} \int_0^L \dot{u}^2 dx = \frac{\mu}{2} \int_0^L \left(\sum_{n=1}^{\infty} \dot{\phi}_n \sqrt{2/L} \sin \frac{n\pi x}{L} \right)^2 dx \\ &= \frac{\mu}{L} \int_0^L \sum_{n=1}^{\infty} \dot{\phi}_n^2 \sin^2 \frac{n\pi x}{L} dx, \end{aligned}$$

since all the product terms disappear because of the orthogonal properties of the normal functions. Thus T_1 becomes reduced to a sum of squares in the generalized velocities and the integration over x leads to the result

$$T_1 = \frac{\mu}{2} \sum_{n=1}^{\infty} \dot{\phi}_n^2. \quad (9)$$

In a similar manner we set up the expression for V , the potential energy. We have

$$V = \frac{T}{2} \int_0^L \left(\frac{\partial u}{\partial x} \right)^2 dx = \frac{T}{2} \int_0^L \left(\sum_{n=1}^{\infty} \phi_n \sqrt{2/L} \frac{n\pi}{L} \cos \frac{n\pi x}{L} \right)^2 dx,$$

which we treat exactly as in the case of the kinetic energy and obtain V as a sum of squares of the generalized coordinates ϕ_n , namely

$$V = \frac{T}{2} \sum_{n=1}^{\infty} \frac{n^2 \pi^2}{L^2} \phi_n^2. \quad (10)$$

Using the Lagrangian equations of motion, we have

$$L_1 = T_1 - V,$$

$$\begin{aligned}\frac{d}{dt}\left(\frac{\partial L_1}{\partial \dot{\phi}_n}\right) &= \mu \ddot{\phi}_n; \\ \frac{\partial L_1}{\partial \phi_n} &= -\left(\frac{n\pi}{L}\right)^2 T \phi_n,\end{aligned}$$

so that the equations of motion become

$$\mu \ddot{\phi}_n + \left(\frac{n\pi}{L}\right)^2 T \phi_n = \Phi_n, \quad n = 1, 2, \dots, \quad (11)$$

where Φ_n is the generalized external force corresponding to the coordinate ϕ_n . Up to this point we have considered only free vibrations for which Φ_n is zero, but now we have generalized our problem to include such things as forced oscillations. We solve the equations above for the case of free vibrations, obtaining

$$\phi_n = A_n e^{i\omega_n t},$$

with

$$\omega_n = \frac{n\pi}{L} \sqrt{\frac{T}{\mu}}$$

so that our expression for u is just the one originally found, in Eq. (3), Chap. XII. It is thus clear that we have essentially used normal coordinates in our first discussion of the vibrating string.

The generalized force Φ_n is defined so that the work done by the external force during a displacement $d\phi_n$ is $\Phi_n d\phi_n$. During a displacement $d\phi_n$, the corresponding displacement of the string is $a_n \sin \frac{n\pi x}{L} d\phi_n = du$, so that if the force acting on a length of string dx at time t is $f dx$, we find for Φ_n ,

$$\Phi_n = \sqrt{2/L} \int_0^L f \sin \frac{n\pi x}{L} dx. \quad (12)$$

In function space, this is evidently simply the component of Φ along the n th axis. An interesting case occurs when the force acts practically at a point $x = a$, such as when a violin string is plucked or bowed. We then write

$$\Phi_n = \sqrt{2/L} \sin \frac{n\pi a}{L} \int_0^L f dx = F \sqrt{2/L} \sin \frac{n\pi a}{L}.$$

This expression brings out the advantage of the concept of generalized force. For example, if a string is struck or bowed at

its center, then $a = L/2$, and $\Phi_n = 0$ when n is an even integer. This means that this force can have no effect on the even overtones and can only affect the odd overtones. If the string is originally at rest, no matter what kind of force is applied at the center, only odd overtones appear in the resultant vibration. No even overtones ever occur, as the normal coordinates are uncoupled and each normal coordinate behaves just as if all the others were absent. These conclusions, immediately obvious from the expression for Φ_n , are not at all obvious if one considers only the usual force acting at a point of the string.

Another case of interest occurs when a periodic force acts on a point $x = a$ of the string. We then have

$$\Phi_n = F \sqrt{2/L} \sin \frac{n\pi a}{L} \cos \omega t$$

and the equation of motion for ϕ_n is very much like the equation of forced motion of a one-dimensional oscillator. The solution of this equation is then

$$\phi_n = A_n \cos (\omega_n t - \epsilon_n) + \frac{F \sqrt{2/L} \sin (n\pi a/L)}{\mu(\omega_n^2 - \omega^2)} \cos \omega t.$$

The first term is the solution of the homogeneous equation and represents the free vibration of this mode, and the second represents the forced vibration indicating all the characteristics of resonance which we have previously studied.

87. The Vibrating String with Friction.—Thus far we have neglected friction forces which must act in real cases. Let us assume that the motion of our string is opposed by a frictional force such that the force on each element of the string is proportional to its velocity. The partial differential equation of the free motion of the string becomes

$$\frac{\partial^2 u}{\partial t^2} + k \frac{\partial u}{\partial t} = \frac{T}{\mu} \frac{\partial^2 u}{\partial x^2}. \quad (13)$$

We can treat this problem rather simply by noting that there is a function G , called the dissipation function, which is one-half the rate at which energy disappears from the system and which has just the same form as the kinetic energy T_1 . In fact, we have

$$G = \frac{1}{2} \mu k \int_0^L \dot{u}^2 dx.$$

One can easily show that the Lagrangian equations when there is a dissipation function become

$$\frac{d}{dt} \left(\frac{\partial L_1}{\partial \dot{q}_i} \right) - \frac{\partial L_1}{\partial q_i} + \frac{\partial G}{\partial \dot{q}_i} = Q_i. \quad (14)$$

According to the special law of friction we have assumed, the dissipation function has the same form as T_1 , so that if we introduce the normal coordinates ϕ_1, ϕ_2, \dots etc., which we found reduced the expressions for T_1 and V to sums of squares, they will also do the same for G , so that we can separate the equations of motion for each coordinate ϕ_n . Proceeding as in the last paragraph, we find

$$G = \frac{\mu k}{2} \sum_{n=1}^{\infty} \dot{\phi}_n^2.$$

The equation for ϕ_n then becomes

$$\ddot{\phi}_n + k\dot{\phi}_n + \omega_n^2 \phi_n = \frac{1}{\mu} \Phi_n \quad (15)$$

which is the same form as in the case of a one-dimensional damped oscillator. From this we see that each of the overtones has the same logarithmic decrement, so that in a free vibration the various overtones maintain their relative amplitudes.

In the case of a forced vibration caused by a periodic force $F \cos \omega t$ acting at the point $x = a$, we have

$$\Phi_n = F \sqrt{2/L} \sin \frac{n\pi a}{L} \cos \omega t,$$

and the steady-state vibrations are given by

$$\phi_n = \frac{F}{\mu} \sqrt{2/L} \sin \frac{n\pi a}{L} \left[\frac{(\omega_n^2 - \omega^2) \cos \omega t + \omega k \sin \omega t}{(\omega_n^2 - \omega^2)^2 + (\omega k)^2} \right]. \quad (16)$$

This is, of course, essentially the same solution we obtained in the discussion of a one-dimensional oscillator.

The particularly simple solutions just obtained depend entirely on the simple form of the law of friction we have assumed. In general, for vibrating systems, the presence of frictional forces does not prevent us from setting up the kinetic and potential energies as a sum of squares. But this transformation will in general not transform the dissipation function G to a sum of

squares. Only in very special cases, such as the law of friction assumed above, does the transformation also reduce G to a sum of squares. The general equation of motion for the coordinate ϕ_1 , for example, will be of the form

$$a_1\ddot{\phi}_1 + c_1\dot{\phi}_1 + \sum_i b_{1i}\phi_i = \Phi_1$$

instead of the simpler form obtained above

$$a_1\ddot{\phi}_1 + b_1\dot{\phi}_1 + c_1\phi_1 = \Phi_1$$

in which we have only ϕ_1 appearing. Thus in the general case of frictional forces there is coupling between the various coordinates so that we have much more complicated types of motion. In Chap. XI, Prob. 3, we had such a case with two coupled circuits with resistance and found that we could get a simple solution only for very small frictional forces.

Problems

1. Write down the Hamiltonian function for a vibrating string, using normal coordinates. Set up Hamilton's equations, and show that they are satisfied for the solution we have found.

2. A sinusoidal force of constant amplitude but adjustable frequency acts on an arbitrary point of a string. The string is in addition damped by a frictional force proportional to the velocity. Discuss the resonance of the string to the force, computing, for example, the total energy of the string as a function of the applied frequency, and showing that the resulting resonance curve goes through maxima corresponding to the various overtone frequencies. Find approximate heights and breadths of the maxima. Neglect the transient vibrations.

3. Prove the orthogonality relations for the normal functions for the weighted string; that is, prove

$$\sin \frac{n\pi x_1}{L} \sin \frac{m\pi x_1}{L} + \cdots + \sin \frac{n\pi x_N}{L} \sin \frac{m\pi x_N}{L} = 0.$$

4. Using the orthogonality relations of Prob. 3, and the analogy of the continuous string, set up a method for finding the amplitudes of the various overtones of the weighted string, in terms of the initial displacements and velocities of the particles.

5. Apply the method of Prob. 4 to the special case of two coupled particles, as taken up in Prob. 1, Chap. XI.

6. Apply Prob. 4 to the case of four particles, as in Prob. 1, Chap. XII.

7. Consider two coupled mechanical vibrating systems, with friction. In general, a dissipative function cannot be set up, and the problem of the motion cannot be solved exactly. Show what relations the frictional forces must satisfy in order to have a dissipative function. Write down the corresponding relations also for the electrical case.

8. What sort of force must be applied to a string in order that the forced motion should be a pure vibration of the n th harmonic?

9. Consider the case of two coupled particles as in Prob. 1, Chap. XI. Show that if equal external forces act on both, the overtone in which they vibrate in opposite directions can never be excited.

10. In the case of the two coupled particles of Prob. 1, Chap. XI, assume that at $t = 0$ both particles are at rest, but that one particle is displaced a distance d , the other not being displaced at all. Find the amplitudes of the two overtones, writing down the formulas for the displacements of each particle as functions of time.

CHAPTER XIV

THE STRING WITH VARIABLE TENSION AND DENSITY

In the last two chapters, we have considered the problem of the vibration of a string of constant density and uniform tension. These results may now be extended for the more general case of variable tension and density. We shall not be able to carry through the results in complete detail; for, as we shall see, we are led to a more complicated differential equation, which we cannot solve in general. But we shall find that the theory of expansion in orthogonal functions, and all the general relations, go through just as with the uniform string, so that we can derive a good deal of information. We shall also develop perturbation methods, which can be used when the tension and density have only small deviations from constancy.

The importance of the problems considered in this chapter arises more from what they suggest than from the specific problems considered. Strings of variable density are of small practical importance. But the string is the simplest case of a vibrating continuum. Waves in three dimensions resemble waves on a string. A string of variable density resembles an optical medium of variable index of refraction, and we meet problems of reflection and refraction. Many three-dimensional problems can actually be reduced to one-dimensional cases, and these are all likely then to take on just the character of our string of variable density. It forms, so to speak, the type for much of our more complicated work. In wave mechanics, for instance, most of our problems reduce to a mathematical form which is identical with that of the present chapter. The perturbation theory we develop in this chapter is one set up originally for use with variable strings, yet it has had most important effects in the development of the quantum theory.

88. Differential Equation for the Variable String.—We set up the differential equation of motion exactly as we have done in Chap. XII. In calculating the resultant force on an element dx of our string we found $\left(T \frac{\partial u}{\partial x}\right)_{x+dx} - \left(T \frac{\partial u}{\partial x}\right)_x$, and this is $\frac{\partial}{\partial x} \left(T \frac{\partial u}{\partial x}\right) dx$,

which reduces as before to $T \frac{\partial^2 u}{\partial x^2} dx$ for constant tension. The remainder of the derivation proceeds as before, and the equation of motion becomes:

$$\frac{\partial}{\partial x} \left(T \frac{\partial u}{\partial x} \right) = \mu \frac{\partial^2 u}{\partial t^2}, \quad (1)$$

where both T and μ are now functions of x . If we assume that u is proportional to a function of x times $e^{i\omega t}$, we find that we get an equation for the function of x alone:

$$\frac{d}{dx} \left(T \frac{du(x)}{dx} \right) + \omega^2 \mu u(x) = 0, \quad (2)$$

where this $u(x)$ is the part of u depending on x .

89. Approximate Solution for Slowly Changing Density and Tension.—The above Eq. (2) is a linear second-order differential equation with variable coefficients, on account of the functions T and μ , which depend on x . We can give no general method of exact solution, except the power series method. To apply that, of course, T and μ must be expressed as power series in x . But it turns out that the solutions of the equation are not very different from sines and cosines of x , and a very useful approximate method of solution is based on this fact, good when the density and tension do not change by a large fraction of themselves in one wave length. This approximate solution is simple, and forms a convenient method for discussing the equation qualitatively. The effect of the variable density and tension comes in two ways: first, the wave length depends on the position, and second the amplitude depends on x . Thus, instead of $A \sin$

$\frac{n\pi x}{L}$, as with the uniform string, the actual solution for the function of x can be at least approximately written in the form $u = A(x) \sin B(x)$. We can see easily the form which B must have for the nonuniform string. For plainly $\frac{B(x_2) - B(x_1)}{2\pi}$ must

measure the number of wave lengths between x_1 and x_2 , on account of the way in which B appears in the sine function. But now if λ is the wave length, regarded as a function of x , dx/λ is just the number of wave lengths in distance dx , so that

the total number between x_1 and x_2 is $\int_{x_1}^{x_2} \frac{dx}{\lambda}$, from which evidently

$B(x) = 2\pi \int dx/\lambda$. Since the wave length can also be written $2\pi/\lambda = \omega\sqrt{\mu/T}$, this is equivalent to $B(x) = \omega \int \sqrt{\mu/T} dx$. It is not hard to show that if we set $A = \frac{\text{constant}}{\sqrt[4]{\mu T}}$, the resulting expression

$$A e^{iB} = \frac{\text{constant}}{\sqrt[4]{\mu T}} e^{i\omega \int \sqrt{\mu/T} dx}, \quad (3)$$

or the corresponding real quantity

$$\frac{\text{constant}}{\sqrt[4]{\mu T}} \cos (\omega \int \sqrt{\mu/T} dx - \alpha)$$

forms an approximate solution of the differential equation.

To prove this equation, we may proceed as follows: we assume the solution

$$u = A e^{i\omega \int \sqrt{\mu/T} dx},$$

where A is an undetermined function of x , and substitute in the differential equation. When the necessary differentiations and substitutions are performed, we obtain a differential equation for A , which may be written, after a little manipulation,

$$\lambda^2 \left(\frac{1}{A} \frac{d^2 A}{dx^2} + \frac{1}{T} \frac{dT}{dx} \frac{1}{A} \frac{dA}{dx} \right) + 4\pi i \lambda \left[\frac{1}{A} \frac{dA}{dx} + \frac{1}{4} \left(\frac{1}{T} \frac{dT}{dx} + \frac{1}{\mu} \frac{d\mu}{dx} \right) \right] = 0, \quad (4)$$

where $\lambda = \frac{2\pi}{\omega} \sqrt{\frac{T}{\mu}}$, the wave length of the disturbance. Now we are assuming that μ , T , and consequently A , do not change by a large fraction of themselves in a wave length. Thus the quantities like $\lambda \frac{1}{A} \frac{dA}{dx}$, measuring the fractional change in A in a wave length, are numbers small compared with 1. Their squares, then, and their rates of change in one wave length, can be neglected, and that means that the first set of terms above, in λ^2 , can be neglected in comparison with the second set, in λ . Considering only the latter terms, we can rewrite the Eq. (4)

$$\begin{aligned} \frac{d \ln A}{dx} + \frac{1}{4} \left(\frac{d \ln T}{dx} + \frac{d \ln \mu}{dx} \right) &= 0, \\ \frac{d \ln (A(\mu T)^{1/4})}{dx} &= 0, \quad A(\mu T)^{1/4} = \text{constant}, \end{aligned}$$

giving the solution we wished to prove.

90. Progressive Waves and Standing Waves.—In the problems of Chap. XII, we noted that there were two sorts of waves possible in a uniform string: progressive waves, and standing waves. The progressive waves traveled along with a velocity v ; an example was $\cos \omega(t - x/v)$, in which the displacement has the same value at all points for which $t - x/v = \text{constant}$, or $x = vt + \text{constant}$, points traveling along with velocity v . Similarly in our general case, we can set up a complex solution

$$\frac{\text{constant}}{\sqrt[4]{\mu T}} e^{i\omega\left(t - \int \frac{dx}{v}\right)},$$

where $v = \sqrt{T/\mu}$. The real part is

$$\frac{\text{constant}}{\sqrt[4]{\mu T}} \cos \omega\left(t - \int \frac{dx}{v}\right),$$

where the equation $t - \int \frac{dx}{v} = \text{constant}$ gives, by differentiation, $dx/dt = v$, verifying that the velocity of propagation of the progressive wave is $v = \sqrt{T/\mu}$, varying from point to point along the string. Thus in the general case we can have a progressive wave along the nonuniform string. We shall see later in the chapter, however, that this is only approximately true for strings with slowly varying density and tension. At a rapid variation of constants, a reflected wave is set up, traveling in the opposite direction, and the superposition of direct and reflected waves eventually produces something more like a standing wave.

An example of a standing wave with a uniform string is $\sin \omega t \sin \omega \frac{x}{v}$, or in the general case $\frac{\text{constant}}{\sqrt[4]{\mu T}} \sin \omega t \sin \omega \int \frac{dx}{v}$.

This is a product of a function of t and a function of x , so that such a wave has nodes, values of x for which the function of x is always zero, so that the vibration always has zero amplitude. We have seen that by combination of two progressive waves we can build up a standing wave; similarly by adding two standing waves we can get a progressive wave, as we see from the relation

$$\cos \omega t \cos \omega \int \frac{dx}{v} + \sin \omega t \sin \omega \int \frac{dx}{v} = \cos \omega\left(t - \int \frac{dx}{v}\right).$$

Thus either sort of wave satisfies the differential equation, and we can add solutions as we always can with homogeneous linear differential equations.

Now suppose a string is held at one point. That means that we must limit ourselves to a particular set of solutions of the differential equation: the standing waves which have a node at that point. Thus in our approximate solution, we must take the space function

$$\frac{\text{constant}}{\sqrt[4]{\mu T}} \sin \omega \int_{x_0}^x \frac{dx}{v},$$

where x_0 is the point where the string is held. Suppose we imagine a semi-infinite string, held at one point, with a wave train of finite length approaching the end. The wave is reflected from the end, travels back, and the superposition of the two trains, in opposite directions, forms the standing wave. This wave will have nodes at definite points on the string. It may have any arbitrary frequency, but the nodes will be differently spaced with different frequencies.

If now the string is held at two points, instead of one, we meet a difficulty: with an arbitrary frequency, the string will not have a node at the second point. We must limit our frequency to one of the discrete set for which there are nodes at both ends. Thus the fact of having the string held at both ends automatically sets up a discrete set of possible frequencies of vibration, the overtones, with a particular form of vibration for each. We let the n th overtone have a wave form represented by $u_n(x)$, an angular frequency ω_n . Thus the whole solution may be written

$$u = \sum_n (A_n \cos \omega_n t + B_n \sin \omega_n t) u_n(x), \quad (5)$$

where the constants A_n and B_n are chosen to satisfy the initial conditions at $t = 0$. If our analytic approximation to the function is good, we have

$$u_n(x) = \frac{\text{constant}}{\sqrt[4]{\mu T}} \sin \omega_n \int_{x_0}^x \frac{dx}{v}, \quad (6)$$

with $v = \sqrt{T/\mu}$. Since the displacement is zero not only at x_0 , but also at the other end x_1 , we must have

$$\omega_n \int_{x_0}^{x_1} \frac{dx}{v} = 2\pi \frac{n}{2}, \quad (7)$$

where n is an integer, which as we readily see equals 1 when there are no nodes between the ends, 2 when there is one node, etc. This leads at once to the condition

$$\omega_n = \frac{n\pi}{\int_{x_0}^{x_1} dx/v} \quad (8)$$

for the angular velocities, which for the uniform string reduces to

$\omega_n = \frac{n\pi}{L} \sqrt{\frac{T}{\mu}}$, where $L = x_1 - x_0$ is the length of the string. If our analytic approximation to the functions u_n is not good, we must simply choose those particular functions for our u_n 's which have nodes at x_1 and x_2 , labeling them in order, the one with $n - 1$ nodes between the ends being called u_n , and then must find the angular frequencies connected with these particular functions. We meet such a case, for example, in some of the problems, where the functions u_n are Bessel's functions, and where we simply must look up the nodes in tables of the roots of Bessel's functions. The particular functions u_n satisfying both differential equation and boundary conditions are called normal functions, or characteristic functions, or wave functions, and the frequencies ω_n are sometimes called characteristic numbers.

91. Orthogonality of Normal Functions.—We can now prove easily, and quite generally, that the normal functions u_n are orthogonal. For this purpose we consider two normal functions u_n and u_m , which are solutions of the differential equation. We then have the identities

$$\frac{d}{dx} \left(T \frac{du_n}{dx} \right) + \omega_n^2 \mu u_n = 0,$$

and

$$\frac{d}{dx} \left(T \frac{du_m}{dx} \right) + \omega_m^2 \mu u_m = 0.$$

We multiply the first equation by u_m , the second by u_n , subtract one from the other, and then integrate over the string, which we assume to extend from $x = 0$ to $x = L$. We thus obtain

$$\int_0^L \left[u_m \frac{d}{dx} \left(T \frac{du_n}{dx} \right) - u_n \frac{d}{dx} \left(T \frac{du_m}{dx} \right) \right] dx = (\omega_m^2 - \omega_n^2) \int_0^L \mu(x) u_n u_m dx.$$

The left side integrated by parts yields immediately

$$\left[T \left(u_m \frac{du_n}{dx} - u_n \frac{du_m}{dx} \right) \right]_0^L - \int_0^L T \left(\frac{du_n}{dx} \frac{du_m}{dx} - \frac{du_m}{dx} \frac{du_n}{dx} \right) dx.$$

The integral obviously vanishes, and the integrated part vanishes since both u_n and u_m are zero for $x = 0$ and $x = L$. In general the integrated part would vanish if either u or du/dx vanished at the boundaries, or if an expression of the form $u + \alpha du/dx$ vanished at each boundary. Thus the right side of the equation above yields us as the analogue of our former orthogonality relation

$$\int_0^L \mu(x) u_n u_m dx = 0, \text{ if } n \neq m, \quad (9)$$

since, when $n = m$, the integral need not vanish to satisfy the original equation. We shall assume the functions to be normalized so that

$$\int_0^L \mu(x) u_n^2 dx = 1. \quad (10)$$

In the previous chapter, where the density μ was independent of x , we could simply omit that factor in the integrals, changing the normalization condition to $\int u^2 dx = 1$, without any error other than a change of a constant factor in the functions u_n . Here, however, the density factor must be kept in. We can see the analogy to the corresponding situation with the two coupled particles. There, if the masses of the particles were m_1, m_2 , and their displacements were y_1, y_2 , we had to set up new quantities x_1, x_2 , equal to $\sqrt{m_1}y_1$ and $\sqrt{m_2}y_2$, respectively. We could give the normalization conditions by stating, for example, that the unit vector along X has unit magnitude. The coordinates of the extremity of this vector, in the notation of Chap. XI, were $x_1 = \alpha', x_2 = \beta'$. Squaring the magnitude of this vector, we had the orthogonality condition

$$x_1^2 + x_2^2 = \alpha'^2 + \beta'^2 = 1.$$

But this is equal to $m_1 y_1^2 + m_2 y_2^2$, where the y 's are the actual displacements. Thus in that case, just as here, we must weight the squares or products of displacements, where they appear in the orthogonality or normalization conditions, with the respective masses. Here the term $\mu(x)dx$ is just the mass of the element dx , so that the analogy is complete.

92. Expansion of an Arbitrary Function Using Normal Functions.—We have seen that we can write our solution

$$u = \sum_n (A_n \cos \omega_n t + B_n \sin \omega_n t) u_n(x).$$

If the initial conditions are $u(x, 0) = f(x)$ and $\frac{\partial u}{\partial t}(x, 0) = F(x)$, where $u(x, t)$ is the function of coordinate and time, we have, substituting in our general solution,

$$\begin{aligned} f(x) &= \sum_n A_n u_n; \\ F(x) &= \sum_n B_n \omega_n u_n, \end{aligned} \quad (11)$$

and we have the general problem of expanding an arbitrary function in a series of normal functions, very much like our previous problem of expanding an arbitrary function in a Fourier series. As before, we shall content ourselves with showing that we can find expressions for the coefficients A_n and B_n which formally satisfy this type of expansion. The remainder of the problem, showing that the series so built up really represents the function and that it converges, will not be taken up here. It is sufficient to say that such proofs can be given.

Let us multiply each of Eqs. (11) on both sides by $\mu(x)u_m$, and integrate from $x = 0$ to $x = L$. We thus have

$$\int_0^L \mu(x) u_m f(x) dx = \sum_n A_n \int_0^L \mu(x) u_m u_n dx$$

and

$$\int_0^L \mu(x) u_m F(x) dx = \sum_n \omega_n B_n \int_0^L \mu(x) u_m u_n dx.$$

On the right side of each of these equations each term for which $m \neq n$ vanishes because of our orthogonality relations. The remaining term contains an integral which has the value unity if the functions u_n are normalized. Thus the whole sum reduces to A_m (or in the second equation to $\omega_m B_m$), and we have found expressions for our coefficients:

$$A_m = \int_0^L \mu(x) f(x) u_m dx,$$

and

$$B_m = \frac{1}{\omega_m} \int_0^L \mu(x) F(x) u_m dx. \quad (12)$$

It is clear that our discussion of the Fourier expansion is but a special case of the general one here discussed. The most convenient point of view to take is that the expression corresponding to the scalar product of two functions $f(x)$ and $\phi(x)$ is

$$\int_0^L \mu(x) f(x) \phi(x) dx.$$

Then clearly our orthogonality and normalization conditions are just what we should expect from our discussion of orthogonal vectors in function space, in the last chapter. The rotation of coordinates in function space again separates variables, as it did in the case of the uniform string; but now the separate normal or characteristic functions are more complicated in form, as we see from the more complicated differential equations they satisfy, though they still vibrate sinusoidally with time. When we carry out an expansion of a function $f(x)$ in terms of the characteristic functions, the coefficients, as with the Fourier expansion, are just the scalar products of the corresponding characteristic functions with the given function, or

$$\int_0^L \mu(x) f(x) u_n dx,$$

as we wrote above.

93. Perturbation Theory.—One approximate method of integrating the differential equation of the nonuniform vibrating string has already been indicated, making use of the resemblance of the actual functions to sines and cosines. An entirely different approximate method, the method of perturbations, is also frequently useful. This is a method which applies if the problem is very nearly a soluble one, the density and tension varying only slightly from their values in the soluble case. The usual application is to an almost uniform string. For simplicity we consider only the case where the tension T is a constant, while the density is a function $\mu(x)$, almost equal to $\mu_0(x)$, for which the problem can be solved. We assume that we know the characteristic functions u_n^0 and frequencies ω_n^0 for the soluble case, satisfying, therefore, the differential equations

$$T \frac{d^2 u_n^0}{dx^2} + \omega_n^{0^2} \mu_0(x) u_n^0 = 0. \quad (13)$$

We now remember that the functions u_n^0 form an orthogonal set, and that any arbitrary function can be expanded in series of such functions. Thus in particular the n th characteristic function u_n of the real problem can be so expanded:

$$u_n = \sum_k A_{nk} u_k^0. \quad (14)$$

We may regard our problem as that of determining the constants A_{nk} . Considered in function space, this problem is very simple.

The functions u_k^0 form one set of orthogonal unit vectors, the u_n 's another, and these equations merely express one set in terms of the other; they are the equations for a rotation of coordinates in function space, from the axes characteristic of the "unperturbed" problem with density μ_0 to the "perturbed" problem with density μ .

The easiest way of getting at the conditions for rotation is simply to substitute u_n in the differential equation which we wish it to satisfy,

$$T \frac{d^2 u_n}{dx^2} + \omega_n^2 \mu u_n = 0.$$

If we do so, and use the differential equations which u_n^0 's satisfy, we have easily

$$\sum_k A_{nk} (\omega_k^2 \mu_0 - \omega_n^2 \mu) u_k^0 = 0.$$

Now we may multiply by an arbitrary u_m^0 , and integrate from 0 to L . Remembering that the u^0 's are orthogonal, the result is

$$\sum_k A_{nk} (\omega_k^2 \mu_0^{mk} - \omega_n^2 \mu_{mk}) = 0, \quad (15)$$

where $\mu_0^{mk} = \int_0^L \mu_0(x) u_m^0 u_k^0 dx = 1$ if $m = k$, 0 if $m \neq k$, and $\mu_{mk} = \int_0^L \mu(x) u_m^0 u_k^0 dx$, a quantity differing from μ_0^{mk} only by small quantities of the order of the deviation between μ and μ_0 . We have here an infinite set of simultaneous homogeneous linear equations (m can take on any value) for the unknown constants A_{nk} . These can be written, for a given n ,

$$\begin{aligned} A_{n1}(\omega_1^2 - \omega_n^2 \mu_{11}) + A_{n2}(-\omega_n^2 \mu_{12}) + A_{n3}(-\omega_n^2 \mu_{13}) + \dots &= 0 \\ A_{n1}(-\omega_n^2 \mu_{21}) + A_{n2}(\omega_2^2 - \omega_n^2 \mu_{22}) + \dots &= 0 \\ A_{n1}(-\omega_n^2 \mu_{31}) + \dots &= 0 \\ \dots &= 0. \end{aligned} \quad (16)$$

In general these will have no solutions; the condition for existence of a solution is that the determinant of coefficients vanish. This forms an equation for ω_n^2 , called a secular or determinantal equation, and just analogous to that which we found with the problem of two coupled vibrations, when we made a rotation of coordinates, and we recognize it as the general type met in

such problems. In this case, the equation has an infinite number of roots, one near each unperturbed frequency.

It is hardly feasible to solve the determinantal equation directly, though it is not hard to make an approximation to it. It is easiest, however, to proceed directly from the linear equations. If the u 's are nearly the same as the u 's, it is plain that we shall have $A_{nk} = 1$ almost, if $n = k$, or $= 0$ almost, if $n \neq k$. The only term in the equations which is large and need be considered is then that for which $n = k$ (so that A_{nk} will be large) and simultaneously $m = k$ (so that μ_{mk}^0 and μ_{mk} will be large). This term gives

$$A_{nn}(\omega_n^{0^2} - \omega_n^2 \mu_{nn}) = 0, \text{ or } \omega_n^2 = \frac{\omega_n^{0^2}}{\mu_{nn}}.$$

If now $\mu = \mu_0 + \mu_1$, where μ_1 is small compared with μ_0 , we have $\mu_{nn} = 1 + \int_0^L \mu_1 u_n^{0^2} dx$, so that, using the first term of a binomial expansion,

$$\omega_n^2 = \omega_n^{0^2} \left(1 - \int_0^L \mu_1 u_n^{0^2} dx \right), \quad (17)$$

correct to the first order of small quantities, but neglecting terms of the order of the square of the integral of μ_1 . It is not hard to get expressions of the same order of accuracy for the A 's.

94. Reflection of Waves from a Discontinuity.—We mentioned earlier that a progressive wave striking a discontinuity of density would be partly reflected, and only partly transmitted. It is easy to solve exactly the problem of propagation of the wave over the discontinuity, and as this is one of the exactly soluble cases of the vibration of the nonuniform string, and is the simplest problem of reflection, it is worth carrying its discussion through. Let us assume two uniform strings of different densities attached to each other and subject to the same tension T . Let the first string have a linear density μ_1 and the second a density μ_2 . We shall take the point of junction as $x = 0$. We thus have different velocities of propagation $v_1 = \sqrt{T/\mu_1}$ and $v_2 = \sqrt{T/\mu_2}$ in the two strings. We may also define an "index of refraction" of one medium with respect to the other as $n = v_1/v_2 = \sqrt{\mu_2/\mu_1}$. At $x = 0$ we must satisfy certain conditions at every instant of time. First, the displacement u must be continuous across the boundary if the strings remain joined together, and secondly, the slope du/dx must also vary continuously across the boundary.

Were the latter condition not fulfilled, we would have the impossible situation of a finite force acting on an infinitesimal piece of the strings at the junction.

Let us consider a harmonic progressive wave in the first string (μ_1) impinging on the junction. In the second string we shall have a wave traveling in the same direction as the impinging wave, but in order to satisfy the boundary conditions, we must assume a reflected wave in the first string. Thus

$$u_1 = Ae^{2\pi i\left(\nu t - \frac{x}{\lambda_1}\right)} + Be^{2\pi i\left(\nu t + \frac{x}{\lambda_1}\right)}$$

and

$$u_2 = Ce^{2\pi i\left(\nu t - \frac{x}{\lambda_2}\right)}.$$

The frequency is a fixed characteristic of the wave, independent of the medium in which the wave is propagated. The wave lengths λ_1 and λ_2 are related by the condition

$$\nu = \frac{v_1}{\lambda_1} = \frac{v_2}{\lambda_2},$$

or

$$\frac{\lambda_1}{\lambda_2} = \sqrt{\frac{\mu_2}{\mu_1}} = n.$$

At the junction, where $x = 0$, we have

$$\begin{aligned}(u_1)_0 &= Ae^{2\pi i\nu t} + Be^{2\pi i\nu t} \\ (u_2)_0 &= Ce^{2\pi i\nu t},\end{aligned}$$

and

$$\begin{aligned}\left(\frac{du_1}{dx}\right)_0 &= -\frac{2\pi i}{\lambda_1}Ae^{2\pi i\nu t} + \frac{2\pi i}{\lambda_1}Be^{2\pi i\nu t} \\ \left(\frac{du_2}{dx}\right)_0 &= -\frac{2\pi i}{\lambda_2}Ce^{2\pi i\nu t}.\end{aligned}$$

Thus the conditions of continuity give

$$A + B = C,$$

and

$$\frac{A}{\lambda_1} - \frac{B}{\lambda_1} = \frac{C}{\lambda_2},$$

whence

$$\frac{B}{A} = \frac{\lambda_2 - \lambda_1}{\lambda_1 + \lambda_2} = -\frac{n - 1}{n + 1},$$

giving the ratio of the amplitude of the reflected to the incident wave. Two limiting cases are interesting: if $\mu_2 = \infty$, so that

the junction is held fast, we have $n = \infty$, $B = -A$, or the wave is entirely reflected, with a change of phase. The other case is $\mu_2 = 0$, the junction is free, and we have $n = 0$, $B = A$, reflection again being complete, but with no change of phase. In both these cases the incident and reflected waves combine to give standing waves.

Problems

1. A heavy uniform flexible chain hangs freely from one end. The chain performs small lateral vibrations. Show that the normal functions are

$$u_n = J_0\left(\frac{2\omega_n}{\sqrt{g}}\sqrt{x}\right), \text{ where } J_0 \text{ represents the Bessel function of order zero;}$$

x is the distance from the bottom of the chain to any point, g the acceleration of gravity and ω_n is the angular frequency of the n th mode of vibration. For a chain 8 feet long, find the periods of the first few modes of vibration (use Jahnke Emde's tables to get the roots of the Bessel functions).

2. One end of a uniform flexible chain of length l is attached to a vertical rod which rotates at a constant angular velocity Ω_0 . Neglect the effect of gravity, so that the chain stands out horizontally under the tension of centrifugal force. Show that the differential equation for small vibrations transverse to the length of the chain is

$$\frac{\Omega_0^2}{2} \frac{d}{dx} \left[(l^2 - x^2) \frac{du}{dx} \right] + \omega^2 u = 0.$$

Introduce the variable $y = x/l$, and solve the resulting equation by the power series method. The boundary conditions are $u(0) = 0$ and u for $y = 1$ must remain finite. Note that the latter condition can only be fulfilled if the series breaks off to form a polynomial. Calculate the first three polynomials and derive a relation for the frequency of the n th mode of vibration. The polynomials so found are the Legendre polynomials of odd order.

3. A string stretched with a uniform tension T , and with a density α/x^2 , is held at the points $x = x_1$ and $x = x_2$. Solve the equation, using the form $y = \sqrt{x}z$, and show that the general solution is

$$u = Ax^{\frac{1}{2}+ik} + Bx^{\frac{1}{2}-ik},$$

where k is defined by $k^2 + \frac{1}{4} = \omega^2 \alpha / T$, and ω is the angular velocity. Show from this that the general form of the normal function is

$$u_n = \sqrt{x} \sin \frac{n\pi}{\ln(x_2/x_1)} \ln(x/x_1), \quad n = 1, 2, 3, \dots,$$

and that

$$\omega_n^2 = \frac{T}{\alpha} \left[\frac{1}{4} + \frac{n^2 \pi^2}{(\ln x_2/x_1)^2} \right].$$

4. Solve the differential equation of Prob. 3 by the approximate method described in this chapter, and show that the solution has the same form as the exact solution. Show that the two solutions coincide in the limit of large α (and also of large x).

5. A progressive wave travels on a uniform string which at $x = 0$ is connected to a string whose density is $\mu = \mu_0 + \alpha x$. This second string is connected to a third at $x = l$ which has the constant density $\mu = \mu_0 + \alpha l$ and the whole is stretched with a uniform tension T . Using the approximate method, find the ratio of the amplitude of the wave transmitted in the third string to the original amplitude of the incident wave in the first string.

6. Consider a string of uniform density μ , length L , but with a tension T which varies slightly from an average tension T_0 . Show with the help of a perturbation calculation that the angular frequency of the n th mode is given approximately by

$$\omega_n^2 = \frac{n^2 \pi^2}{L^2} \frac{T_0}{\mu} \left(1 - \frac{2}{n \pi T_0} \int_0^L \frac{dT}{dx} \cos \frac{n \pi x}{L} \sin \frac{n \pi x}{L} dx \right).$$

7. A uniform string of density μ_0 , tension T , has a small load m placed at $x = a$. Show that the frequency of the n th mode of vibration is approximately given by

$$\omega_n^2 = \frac{n^2 \pi^2}{L^2} \frac{T}{\mu_0} \left(1 - \frac{2m}{\mu_0 L} \sin^2 \frac{n \pi a}{L} \right).$$

Show that the effect of the additional load vanishes if it is placed at a node, and is biggest when at an antinode.

8. Show that the differential equation of Bessel's function J_m is the same as that for a string of tension $T = x$, $\mu \omega^2 = x - m^2/x$. Using the approximate method developed for the vibration problem, show that approximately

$$J_m(x) = \frac{\text{constant}}{\sqrt[4]{x^2 - m^2}} \cos \left(\int \sqrt{1 - m^2/x^2} dx - \alpha \right),$$

where $x > m$.

9. Using the approximation of Prob. 8 for J_0 and J_1 , compute the approximation functions for a number of values of x , and show by a table of values how well these agree with the correct functions. Choose the arbitrary amplitude and phase factors to make the functions agree with the values of J_0 and J_1 in the tables, for example making the zeros agree by adjusting α , and the maxima by adjusting the amplitude, taking such values as to get the best agreement possible for large x 's.

10. Derive the differential Eq. (4) for A , in the approximate solution

$$u = A e^{i \omega \int \sqrt{\mu/T} dx}.$$

CHAPTER XV

THE VIBRATING MEMBRANE

The problem of a vibrating membrane is very little more difficult in principle than the string. Let us take two coordinates, x and y , in the plane of the membrane, writing u for the displacement at right angles to the plane, so that we wish a relation $u = u(x, y, t)$. Consider a small element of the membrane, bounded by dx and dy . Let the mass per unit area be μ , so that the mass of the element is $\mu dx dy$. Then its mass, times acceleration normal to the membrane, is $\mu dx dy \partial^2 u / \partial t^2$. This is equal to the force arising from the tension. Let the tension be T . That is, if we cut the membrane along any line, the material on one side of the cut exerts a force on the material on the other, normal to the cut and equal to T for each unit of length of the cut. We assume that T is constant over the membrane. If the membrane were plane, the tension on its opposite edges would cancel, and we should have no resultant force. If it is curved, however, we may proceed as follows. Along the edge at $x + dx$, the tension is at right angles to the y axis, almost along the x axis, but with a small component along the u direction, equal approximately to $T \left(\frac{\partial u}{\partial x} \right)_{x+dx}$ per unit of length, or this times dy for the actual length dy . Similarly along the edge at x the component is $-T \left(\frac{\partial u}{\partial x} \right)_x dy$, so that the sum is approximately $T \partial^2 u / \partial x^2 dx dy$. The forces acting along the edges at y and $y + dy$ similarly add to $T \partial^2 u / \partial y^2 dx dy$, and the total force, the sum of these, is $T (\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2) dx dy$. Thus the differential equation, dividing by $dx dy$, is

$$\mu \frac{\partial^2 u}{\partial t^2} = T \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right). \quad (1)$$

95. Boundary Conditions on the Rectangular Membrane.—A membrane is ordinarily held fast around a certain curve. In this way one can get a great variety of problems, by taking different curves. The two simplest are the rectangular mem-

brane, and the circular membrane, or ordinary drum, and in the present section we consider the rectangular case, assuming the membrane to be held at $x = 0$, $x = X$, $y = 0$, $y = Y$. We solve first by the exponential method, assuming

$$u = e^{i(\omega t + kx + ly)}.$$

Then the differential equation becomes $-\mu\omega^2 = -T(k^2 + l^2)$, $\omega = \sqrt{T(k^2 + l^2)/\mu}$, giving the angular velocity of the vibration in terms of the quantities k and l . Instead of the exponential solution we can equally well use sines or cosines. For example, with a given ω , k , and l , we can take

$$\begin{aligned} u &= e^{i\omega t}(e^{ikx+ily} - e^{-ikx+ily} - e^{ikx-ily} + e^{-ikx-ily}) \\ &= e^{i\omega t}(2i \sin kx)(e^{ily} - e^{-ily}) \\ &= -4e^{i\omega t} \sin kx \sin ly. \end{aligned}$$

As a matter of fact, this solution with sines is the one we want, since it reduces to zero when $x = 0$ and $y = 0$. To apply the condition when $x = X$ and $y = Y$, we must make the sines zero at these points, or must have $\sin kX = 0$, $\sin lY = 0$, or $k = n\pi/X$, $l = m\pi/Y$, where n , m are integers. In terms of these constants, we can then write

$$\omega = \pi\sqrt{\frac{T}{\mu}\left(\frac{n^2}{X^2} + \frac{m^2}{Y^2}\right)}, \quad (2)$$

so that instead of having overtones whose frequencies are integral multiples of a fundamental, the frequencies are given by a much more complicated relation. There is one interesting result of this. Pleasing musical notes depend on having the frequencies of the overtones related in simple ways to the fundamental, so that they sound well together, as with a vibrating string. In a membrane or drum, in which these relations do not hold, the sound is far less musical than with a string. This suggests other cases, which do not exactly fall within the category of the present chapter. For example, a vibrating bell acts as a two-dimensional vibrating system, a little like a membrane, and has complicated overtones which in general are not harmonics. But it has been found by trial that if bells are made in their conventional shape, overtones are so adjusted that the loud ones are actually in tune with each other, though a slight change of shape would destroy the quality.

96. The Nodes in a Vibrating Membrane.—If the membrane is vibrating with one overtone, the amplitude will be zero along certain lines, which will stay at rest. These nodal lines form a rectangular network, coming when $nx/X = 1, 2, \dots, n-1$, and for $my/Y = 1, 2, \dots, m-1$. At any instant, if the membrane is displaced upward in one rectangle, it will be displaced downward in all adjacent rectangles. Such a nodal arrangement is characteristic of all sorts of standing wave problems

97. Initial Conditions.—At $t = 0$, we may wish to fix the shape and velocity of our membrane, obtaining initial conditions of the sort found with the string, and leading as before to Fourier series. For example, suppose the initial velocity is zero, the initial displacement a function $f(x, y)$. Then we must have

$$u = \sum_{n,m} A_{nm} \cos \omega_{nm} t \sin \frac{n\pi x}{X} \sin \frac{m\pi y}{Y}, \quad (3)$$

where ω_{nm} is given in Eq. (2), and where we have

$$f(x, y) = \sum_{n,m} A_{nm} \sin \frac{n\pi x}{X} \sin \frac{m\pi y}{Y}. \quad (4)$$

To find the coefficients A , the amplitudes of the various overtones necessary to satisfy the conditions, we must expand the function $f(x, y)$ in a series of products of sines—a double Fourier series, as it is called. As in the last chapter, we assume that the expansion can be carried through, and ask only for the values of the coefficients. Multiplying both sides of the equation by $\sin \frac{n'\pi x}{X} \sin \frac{m'\pi y}{Y}$, where n', m' are definite integers, we integrate with respect to x from 0 to X , and with respect to y from 0 to Y . We find as before that $\int_0^X \sin \frac{n\pi x}{X} \sin \frac{n'\pi x}{X} dx$ is zero unless $n = n'$, and is $X/2$ if $n = n'$. Thus the final result is

$$A_{n'm'} = \frac{4}{XY} \int_0^Y \int_0^X f(x, y) \sin \frac{n'\pi x}{X} \sin \frac{m'\pi y}{Y} dx dy. \quad (5)$$

It is worth noting that this is the first time we have had to use a double integral. If $f(x, y)$ is a complicated function of the coordinates, it can, of course, be a very difficult problem actually to evaluate the integral.

98. The Method of Separation of Variables.—To solve our differential equation, we may adopt a slightly different method, called the method of separation of variables, which does not directly depend on the use of exponentials. It is a method for reducing the partial differential equation to a set of ordinary equations, and we shall find it very useful. In fact, it is so valuable that practically the only partial differential equations which can be solved at all are those for which this method can be used.

We wish to solve $\frac{\partial^2 u}{\partial t^2} = \frac{T}{\mu} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$. Suppose we try to find a solution u which is the product of a function of x , a function of y , and a function of t ; say $u = P(x)Q(y)R(t)$, where P is a function of x to be determined, and so on. Of course, it is not obvious that one can find such a solution, but our experience would lead us to try it. If we substitute, we have, for example, $\partial u / \partial t = PQ \, dR / dt$, and so on. If we denote dR / dt by R' , with corresponding notation, we then have $PQ \, R'' = (T/\mu) (P'' QR + P Q'' R)$. Next we divide by PQR , obtaining

$$\frac{R''}{R} = \frac{T}{\mu} \left(\frac{P''}{P} + \frac{Q''}{Q} \right). \quad (6)$$

We now make the step characteristic of the method of separation of variables: we observe that the function R''/R on the left of Eq. (6) is a function of t alone, the quantity on the right a function of x and y alone. The equation then states that a certain function of t equals a function of x and y , whatever x , y , and t may be. But this is clearly impossible in general. If, for example, we keep x and y constant, and vary t , the left side would change, the right remaining constant, and the equation would not be satisfied. The only exception, as this example shows, is if the left side is a constant, independent of t , and similarly if the right side is a constant, independent of x and y . Let us then impose these conditions, letting the constant be $-\omega^2$ (an arbitrary constant so far, but later to be identified with our other ω). We have then two equations,

$$\frac{R''}{R} = -\omega^2,$$

or

$$R'' + \omega^2 R = 0,$$

and

$$\frac{T}{\mu} \left(\frac{P''}{P} + \frac{Q''}{Q} \right) = -\omega^2. \quad (7)$$

Taking the latter equation, we may again separate. We write it

$$\frac{P''}{P} = -\frac{Q''}{Q} - \omega^2 \frac{\mu}{T}. \quad (8)$$

The left side is a function of x , the right side of y , and by the same argument each is a constant, say $-k^2$. Then we have $P'' + k^2P = 0$, and $-k^2 = -(Q''/Q) - \omega^2(\mu/T)$. We can rewrite this last $Q''/Q = -l^2$, where

$$-l^2 = k^2 - \omega^2 \frac{\mu}{T}, \text{ or } \omega^2 = \frac{T}{\mu}(k^2 + l^2), \quad (9)$$

and it becomes $Q'' + l^2Q = 0$. We now have three ordinary differential equations for P , Q , and R , whose solutions are evidently

$$P = e^{ikx} (\text{or } e^{-ikx}, \text{ or } \sin kx, \text{ or } \cos kx), \\ Q = e^{ily}, R = e^{i\omega t},$$

so that the final solution is as we found before, with the same relation between ω , k , and l .

99. The Circular Membrane.—The differential equation for the circular membrane is the same as for the rectangular one, but the boundary condition is different: the displacement u is always zero on a circle of radius ρ about the origin. To solve the problem, the simplest method is to introduce polar coordinates, r , θ ; for then the boundary condition is that $u = 0$ when $r = \rho$, which is a condition easy to apply. Let us then write our equation in polar coordinates. Before doing it, we shall give the conventional names of the equations and symbols we meet. Our equation, which is often written

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} = 0, \quad (10)$$

where $v = \sqrt{T/\mu}$ is the velocity of the wave, is called the wave equation, for u represents waves, either progressive or standing. The special case $\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = 0$, where u is independent of t , is called Laplace's equation. And the expression $\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2$, which we have already seen can be written in vector notation $\nabla^2 u$, is called the Laplacian of u . Our present problem is to find the Laplacian in polar coordinates.

100. The Laplacian in Polar Coordinates.—Let us introduce r and θ by the equations $x = r \cos \theta$, $y = r \sin \theta$, $r = \sqrt{x^2 + y^2}$,

$\theta = \tan^{-1} \frac{y}{x}$, so that $\frac{\partial r}{\partial x} = \frac{x}{r}$, $\frac{\partial r}{\partial y} = \frac{y}{r}$, $\frac{\partial \theta}{\partial x} = \frac{-y}{r^2}$, $\frac{\partial \theta}{\partial y} = \frac{x}{r^2}$, and $\frac{\partial^2 r}{\partial x^2} = \frac{1}{r} - \frac{x^2}{r^3}$, $\frac{\partial^2 r}{\partial y^2} = \frac{1}{r} - \frac{y^2}{r^3}$, $\frac{\partial^2 \theta}{\partial x^2} = \frac{2xy}{r^4}$, $\frac{\partial^2 \theta}{\partial y^2} = \frac{-2xy}{r^4}$. Then we have

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial u}{\partial \theta} \frac{\partial \theta}{\partial x}.$$

If we apply this process again, we find without difficulty

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial r^2} \left(\frac{\partial r}{\partial x} \right)^2 + 2 \frac{\partial^2 u}{\partial r \partial \theta} \left(\frac{\partial r}{\partial x} \frac{\partial \theta}{\partial x} \right) + \frac{\partial^2 u}{\partial \theta^2} \left(\frac{\partial \theta}{\partial x} \right)^2 + \frac{\partial u}{\partial r} \frac{\partial^2 r}{\partial x^2} + \frac{\partial u}{\partial \theta} \frac{\partial^2 \theta}{\partial x^2}.$$

Proceeding similarly with y , and adding, we have

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} &= \frac{\partial^2 u}{\partial r^2} \left[\left(\frac{\partial r}{\partial x} \right)^2 + \left(\frac{\partial r}{\partial y} \right)^2 \right] + 2 \frac{\partial^2 u}{\partial r \partial \theta} \left(\frac{\partial r}{\partial x} \frac{\partial \theta}{\partial x} + \frac{\partial r}{\partial y} \frac{\partial \theta}{\partial y} \right) + \\ &\quad \left(\frac{\partial^2 u}{\partial \theta^2} \right) \left[\left(\frac{\partial \theta}{\partial x} \right)^2 + \left(\frac{\partial \theta}{\partial y} \right)^2 \right] + \frac{\partial u}{\partial r} \left(\frac{\partial^2 r}{\partial x^2} + \frac{\partial^2 r}{\partial y^2} \right) + \frac{\partial u}{\partial \theta} \left(\frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} \right). \end{aligned}$$

Substituting, this becomes

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{1}{r} \frac{\partial u}{\partial r},$$

which can also be written

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2}. \quad (11)$$

This is the expression for the Laplacian in polar coordinates.

101. Solution of the Differential Equation by Separation.—Our differential equation is now

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2}. \quad (12)$$

Let us solve by separation of variables, assuming $u = R(r)\Theta(\theta)T(t)$. Then, substituting, and dividing by $R\Theta T$, the result is

$$\frac{1}{R} \frac{1}{r} \frac{d}{dr} \left(r \frac{dR}{dr} \right) + \frac{1}{r^2} \frac{1}{\Theta} \frac{d^2 \Theta}{d\theta^2} = \frac{1}{v^2} \frac{1}{T} \frac{d^2 T}{dt^2}. \quad (13)$$

The problem is separated: the left side depends only on r and θ , the right on t . Each must then be a constant, which we shall call $-\omega^2/v^2$, giving $d^2 T/dt^2 + \omega^2 T = 0$, $T = A \cos \omega t + B \sin \omega t$,

and

$$\frac{1}{R} \frac{1}{r} \frac{d}{dr} \left(r \frac{dR}{dr} \right) + \frac{1}{r^2} \frac{1}{\Theta} \frac{d^2 \Theta}{d\theta^2} = -\frac{\omega^2}{v^2}.$$

We multiply by r^2 , and transfer the first term to the right, obtaining

$$\frac{1}{\Theta} \frac{d^2 \Theta}{d\theta^2} = -r^2 \left[\frac{1}{R} \frac{1}{r} \frac{d}{dr} \left(r \frac{dR}{dr} \right) + \frac{\omega^2}{v^2} \right].$$

Again the variables are separated, the left side depending only on θ , the right on r . Let each equal $-m^2$. Then $d^2\Theta/d\theta^2 + m^2\Theta = 0$, $\Theta = C \cos m\theta + D \sin m\theta$, and the equation for r can be immediately changed to

$$\frac{1}{r} \frac{d}{dr} \left(r \frac{dR}{dr} \right) + \left(\frac{\omega^2}{v^2} - \frac{m^2}{r^2} \right) R = 0. \quad (14)$$

This is just like Bessel's equation (see Prob. 13, Chap. II), except that it has the constant ω^2/v^2 in place of 1. A simple change of variables removes this discrepancy, however. Let $x = \omega r/v$. Then the equation becomes

$$\frac{\omega^2}{v^2} \frac{1}{x} \frac{d}{dx} \left(x \frac{dR}{dx} \right) + \left(\frac{\omega^2}{v^2} - \frac{m^2}{x^2} \right) R = 0;$$

or cancelling ω^2/v^2 , it is exactly Bessel's equation

$$\frac{1}{x} \frac{d}{dx} \left(x \frac{dR}{dx} \right) + \left(1 - \frac{m^2}{x^2} \right) R = 0. \quad (15)$$

The solution is then $R = \text{constant} \times J_m(x)$, a Bessel's function of the m th order, whose expansion in power series we have already considered, for integral values of m , and for which we have found an approximation in the preceding chapter (Chap. XIV, Prob. 8). We shall see in the next section that only integral m 's must be used in the present problem.

102. Boundary Conditions.—Consider in the first place the solution for θ . At a given point of the membrane, the value of θ is determined, but not in a single-valued way. Thus if the point corresponds to $\theta = 47$ deg., it would equally well correspond to 47 deg. + 360 deg., or 47 deg. + 720 deg., etc. Now Θ must surely have a definite value at each point of the membrane. Thus it must have the same value for θ , $\theta + 2\pi$, $\theta + 4\pi$, etc. In other words, Θ is periodic in θ with period 2π . But this is true

if, and only if, m is an integer. Hence our first condition, necessary to make the function single valued, is that m be an integer.

Next consider the solution for r : $R = J_m(\omega r/v)$, where now m is an integer. At the edge of the membrane, $u = 0$, which means that $R = 0$, or $J_m(\omega \rho/v) = 0$. Now $J_m(x)$ is zero only for certain definite values of x , say $x = x_1, x_2, x_3, \dots$. From the properties of Bessel's functions, we have seen that there are an infinite number of such roots. Thus, to satisfy our boundary conditions, we must let $\omega \rho/v = x_1, x_2, \dots$. The only adjustable quantity is ω , so that it must be determined

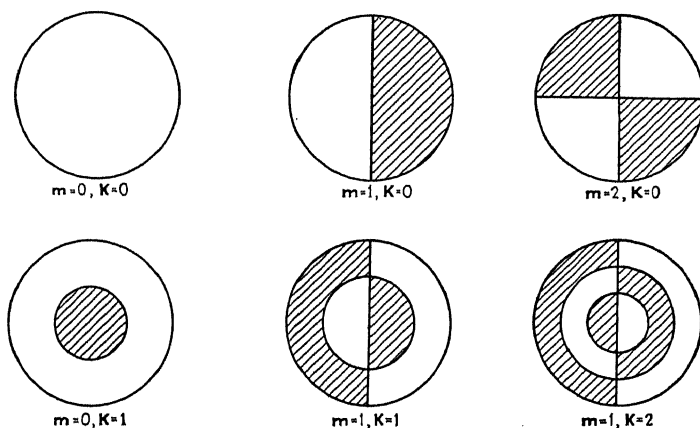


FIG. 22.—Nodes of a circular membrane. Shaded segments are displaced in opposite phase to unshaded.

by one or another of the values $\omega = vx_1/\rho, vx_2/\rho, \dots$. Suppose in particular that $\omega = v x_k/\rho$, determined by the k th root of J_m . Then we should properly label it ω_{mk} , since it depends on both these indices. We have thus determined our solution completely, except for the remaining arbitrary constants. These can be easily expressed in the following form:

$$u = (A \cos \omega_{mk}t + B \sin \omega_{mk}t) \cos (m\theta - \alpha_{mk}) J_m(\omega_{mk}r/v).$$

This is a particular solution. The general solution is the sum of such terms, taken over all m 's and k 's.

103. Physical Nature of the Solution.—A single term corresponds to a single standing wave. Its nodes are concentric circles, values of r for which $J_m(\omega_{mk}r/v)$ is zero, of which of course the boundary is one; and radii, determined by $\cos (m\theta - \alpha) = 0$, as in Fig. 22. It is readily seen that there are m radial

nodes, k circular nodes without counting the boundary. The arbitrary constant α_{mk} determines the angles at which the radial nodes are; changing it simply rotates the whole nodal pattern. The constants A and B determine the amplitude and phase of the disturbance as a function of time. We may, if we choose, consider that there are two separate waves possible for each frequency, $\cos m\theta J_m$ and $\sin m\theta J_m$. Such a case is called degenerate; we shall see in a problem that the same thing is true of the square membrane. In a degenerate case, with two or more possible vibrations of the same frequency, it is plain that any linear combination of these vibrations gives a possible vibration of this same frequency. As with the rectangular membrane, the set of frequencies ω_{mk} does not form a simple set of overtones with pitches in harmonic relation to each other.

104. Initial Condition at $t = 0$.—Suppose we know that at $t = 0$, the displacement of the membrane is given by $F(r, \theta)$, and the velocity by $G(r, \theta)$. Now we can write the whole solution, in a slightly more general way than before,

$$u = \sum_{m,k} [(A_{mk} \cos \omega_{mk}t + B_{mk} \sin \omega_{mk}t) \cos m\theta + (C_{mk} \cos \omega_{mk}t + D_{mk} \sin \omega_{mk}t) \sin m\theta] J_m\left(\frac{\omega_{mk}r}{v}\right). \quad (16)$$

Thus, writing displacement and velocity at $t = 0$, we have

$$\begin{aligned} F(r, \theta) &= \sum_{m,k} (A_{mk} \cos m\theta + C_{mk} \sin m\theta) J_m\left(\frac{\omega_{mk}r}{v}\right) \\ G(r, \theta) &= \sum_{m,k} \left[\omega_{mk} (B_{mk} \cos m\theta + D_{mk} \sin m\theta) J_m\left(\frac{\omega_{mk}r}{v}\right) \right]. \end{aligned} \quad (17)$$

The A 's, B 's, C 's, D 's must be chosen to fit these conditions. Both conditions are of the same sort. They require us to find the coefficients for expanding a function of r and θ in series of products of sines and cosines and Bessel's functions. Now it proves to be true that both the sines or cosines and the Bessel's functions are orthogonal, and as a result of this we can make the expansions we desire in the usual way, as with Fourier series. Let us take the first equation, multiply by $\cos n\theta J_n(\omega_n r/v)$, and integrate over the area of the drum. That is, we integrate with respect to r from 0 to ρ , and with respect to θ from 0 to 2π , and the element of area is $rdrd\theta$. Then we have

$$\begin{aligned} \int_0^\rho \int_0^{2\pi} F(r, \theta) \cos n\theta J_n\left(\frac{\omega_n l r}{v}\right) r dr d\theta \\ = \sum_{mk} \int_0^{2\pi} (A_{mk} \cos m\theta + C_{mk} \sin m\theta) \cos n\theta d\theta \\ \int_0^\rho r J_m\left(\frac{\omega_{mk} r}{v}\right) J_n\left(\frac{\omega_n l r}{v}\right) dr. \end{aligned}$$

By the orthogonal property of the sine and cosine, the right side is zero unless $m = n$, giving $\sum_k \pi A_{nk} \int_0^\rho r J_n\left(\frac{\omega_{nk} r}{v}\right) J_n\left(\frac{\omega_n l r}{v}\right) dr$.

But now we shall prove in the next section that the J 's are orthogonal in the sense that $\int_0^\rho r J_n\left(\frac{\omega_{nk} r}{v}\right) J_n\left(\frac{\omega_{nl} r}{v}\right) dr = 0$, if $k \neq l$. Using this fact, our sum reduces to the single term

$$\pi A_{nl} \int_0^\rho r J_n^2\left(\frac{\omega_n l r}{v}\right) dr.$$

If the last integral, which could be easily computed if we knew the properties of Bessel's functions better, were denoted by c_{nl} , then we should have

$$A_{nl} = \frac{1}{\pi c_{nl}} \int_0^\rho \int_0^{2\pi} F(r, \theta) \cos n\theta J_n\left(\frac{\omega_n l r}{v}\right) r dr d\theta, \quad (18)$$

determining the coefficients A in terms of a single integral. Similarly we could get formulas for the B 's, C 's, D 's. Of course, in an actual case, these integrals might be very difficult to compute, but nevertheless we have a general solution of our problem.

105. Proof of Orthogonality of the J 's.—We can prove the orthogonality of the J 's directly from the differential equation, as was done in the last chapter for the nonuniform vibrating string. We wish to prove that

$$\int_0^\rho r J_n\left(\frac{\omega_n l r}{v}\right) J_n\left(\frac{\omega_{nk} r}{v}\right) dr = 0, \text{ if } l \neq k. \quad (19)$$

Now we have

$$\begin{aligned} \frac{1}{r} \frac{d}{dr} \left[r \frac{dJ_n(\omega_n l r/v)}{dr} \right] + \left(\frac{\omega_n l^2}{v^2} - \frac{n^2}{r^2} \right) J_n\left(\frac{\omega_n l r}{v}\right) &= 0, \\ \frac{1}{r} \frac{d}{dr} \left[r \frac{dJ_n(\omega_{nk} r/v)}{dr} \right] + \left(\frac{\omega_{nk}^2}{v^2} - \frac{n^2}{r^2} \right) J_n\left(\frac{\omega_{nk} r}{v}\right) &= 0. \end{aligned}$$

Multiply the first by $r J_n(\omega_{nk}r/v)$, the second by $r J_n(\omega_{nl}r/v)$, subtract, and integrate from 0 to ρ . The result is

$$\begin{aligned} \int_0^\rho \left\{ J_n\left(\frac{\omega_{nk}r}{v}\right) \frac{d}{dr} \left[r \frac{dJ_n(\omega_{nl}r/v)}{dr} \right] - J_n\left(\frac{\omega_{nl}r}{v}\right) \frac{d}{dr} \left[r \frac{dJ_n(\omega_{nk}r/v)}{dr} \right] \right\} dr \\ = \left(\frac{\omega_{nk}^2 - \omega_{nl}^2}{v^2} \right) \int_0^\rho r J_n\left(\frac{\omega_{nl}r}{v}\right) J_n\left(\frac{\omega_{nk}r}{v}\right) dr. \end{aligned}$$

Just as in the last chapter, the left side can be shown to be zero, by integrating by parts. Then the right side must be zero, and either $\omega_{nk}^2 - \omega_{nl}^2$ is zero, which is not true unless k and l refer to the same overtone, or $\int_0^\rho r J_n(\omega_{nl}r/v) J_n(\omega_{nk}r/v) = 0$, which we wished to prove. The orthogonality is not quite of the form discussed in the last chapter, for the differential equation is of slightly different form, the quantity $(\omega^2/v^2 - n^2/r^2) r$ appearing in place of $\omega^2\mu$, so that the final result is not just like integrating μ times the product of the functions to get zero.

Problems

1. A rectangular drum is 20 by 40 cm., its whole mass is 100 gm., the total pull on the faces 50 and 100 kg., respectively. Find the frequencies, in cycles per second, of the five lowest modes of vibration, and sketch the nodes for each.

2. The special case of degeneracy arises when a rectangular membrane is square. Then the two modes of vibration $e^{i\omega t} \sin(n\pi x/X) \sin(m\pi y/X)$ and $e^{i\omega t} \sin(m\pi x/X) \sin(n\pi y/X)$ have the same frequency (where we let $X = Y$). Thus any linear combination of these is a solution, again with this frequency. Consider the combinations

$$e^{i\omega t} \left(A \sin \frac{n\pi x}{X} \sin \frac{m\pi y}{X} + B \sin \frac{m\pi x}{X} \sin \frac{n\pi y}{X} \right).$$

Work out the nodes in the case $n = 1, m = 2$, for (1) $B = A$; (2) $B = -A$; (3) $B = 2A$.

3. A rectangular membrane is struck at its center, starting from rest, in such a way that at $t = 0$ a small rectangular region about the center may be considered to have a velocity v , and the rest has no velocity. Find the amplitudes of the various overtones.

4. Imagine n and m plotted as two rectangular coordinates. Show that a curve of constant ω , plotted in these coordinates, is an ellipse. Each integral value of n and m corresponds to an overtone, so that if we draw the point corresponding to each overtone, the number of points within such an ellipse gives the number of overtones with angular velocity less than ω . Note that the number of such points per unit area of the plane is just one, and so find an approximate formula, using the area of the ellipse, for the number of overtones of frequency less than ω , and also for the number

between ω and $\omega + d\omega$. Check up this approximation by the exact values of Prob. 1.

5. In the circular membrane, suppose that $m = 0$, and that k is very large, so that there are many circular nodes. Consider a small region near the edge of the membrane. The few nodes in this neighborhood will be almost straight lines, as if we were near the edge of a rectangular membrane. Find the asymptotic wave length, using the fact that $J_m(x)$ approaches $\cos(x - a)$ at large x , and show that the wave length is connected with the velocity and frequency in the usual manner.

6. Set up the wave equation in three-dimensional spherical coordinates, in which $x = r \sin \theta \cos \phi$, $y = r \sin \theta \sin \phi$, $z = r \cos \theta$. Show that it is

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2} = \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2}.$$

7. Separate variables in the preceding equation. Show that the function of ϕ is $\sin m\phi$ or $\cos m\phi$, where m is an integer. Show that the equations for r and θ are respectively

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) + \left(\frac{\omega^2}{v^2} - \frac{C}{r^2} \right) R = 0,$$

where ω , C are constants;

$$\frac{1}{\sin \theta} \frac{d}{d\theta} \left(\sin \theta \frac{d\Theta}{d\theta} \right) + \left(C - \frac{m^2}{\sin^2 \theta} \right) \Theta = 0.$$

8. The equation for θ in Prob. 7 is called Legendre's equation. Let $\Theta = \sin^m \theta F(\cos \theta)$. Find the differential equation for F , solving in power series in $\cos \theta$, and show that the series breaks off if $C = l(l+1)$, where l is an integer. The resulting functions are called $P_l^m(\cos \theta)$, and are known as associated Legendre polynomials. Compute the first few Legendre functions.

9. In the equation for r in Prob. 7, prove that $R = \frac{J_{l+\frac{1}{2}}(x)}{\sqrt{x}}$, where $x = \omega r/v$.

10. Prove that two functions u_n and u_m , satisfying differential equations of the form

$$\frac{d}{dx} \left[T(x) \frac{du_n}{dx} \right] + [\mu(x)\omega_n^2 - f(x)]u_n = 0,$$

with different ω_n 's, but chosen so that both u_n and u_m are zero at $x = 0$ and $x = L$, satisfy the orthogonality condition $\int_0^L \mu(x) u_n(x) u_m(x) dx = 0$.

CHAPTER XVI

STRESSES, STRAINS, AND VIBRATIONS OF AN ELASTIC SOLID

In the preceding chapters, we have been treating the vibrations of elastic strings and membranes, one- and two-dimensional bodies, and now we pass to the three-dimensional case, or the elastic solid. Of course, the strings and membranes were really elastic solids, of particular shapes. But there are several ways in which we must give a more general treatment than we have previously done. First, in the strings and membranes, the rigidity of the material itself was not great enough to affect the vibration, whereas in the problems we now take up this rigidity, or the elastic properties of the material in general, will be important. Thus we may imagine all gradations of the problem of a stretched wire, from the limiting case of a very thin long wire under large tension, when our previous theory is applicable, down to a short thick bar under small tension or even with no tension at all, when the restoring force on a particle, far from coming from the tension on the ends, comes from the distortion of the bar itself. Secondly, with the strings and membranes, we considered only transverse vibrations, while here we discuss longitudinal vibrations as well. Of course, strings can vibrate longitudinally, but we have so far neglected this phase of their motion. Thirdly, a very important part of the problems of strings and membranes has arisen from the fact that they were limited in space, the membranes being very thin pieces of material, the strings thin in two dimensions. But while some of the problems of the present chapter have this property, we shall also consider vibrations and waves in extended media going, in the limiting case, to infinity in all dimensions, as sound waves in an infinite gas or solid. It is these sound waves which show the best analogy to our one- and two-dimensional-wave equations.

106. Stresses, Body and Surface Forces.—The first step in discussing the vibrations of an elastic solid, as with the string and

membrane, is to find the force acting on an infinitesimal volume element, and to set this equal to mass times acceleration. The forces may be divided into two classes: (1) volume or body forces, such as gravity, which act on each volume element of the body, and which for the present we neglect, since we shall not use them in our applications; and (2) surface forces, with which neighboring parts of the medium act on each other, and which are transmitted across surfaces, or the forces transmitted across the bounding surface of the whole body. The tensions which we have met with string and membrane are examples of such forces, or pressures in a gas, or shearing forces in a twisted rod. To specify such a force, we imagine a surface element dA to be drawn somewhere in the body, with a normal n . The material on either side of dA exerts a force on the material on the other side; thus this force is a push normal to the surface if there is a pressure in the body, it is a tension if that is the form of stress, or it may be a shearing force. The force exerted by the material on one side, on the material on the second side, and the other force exerted by the material on the second side back on the first side, are action and reaction, and are equal and opposite, so that one always has an ambiguity of sign in dealing with these forces, or as we call them stresses. We adopt the following convention: We imagine dA to be part of the surface bounding a volume, and n to be the outer normal. Then the force we deal with is the force exerted by the outside on the material inside the volume, over dA . Now this force will be a vector, and proportional to dA ; we call its x , y , and z components $X_n dA$, $Y_n dA$, $Z_n dA$, respectively. The capital letters indicate the force components, and the subscript n denotes not a component but the direction of the surface normal.

The properties of a stress can be completely specified if we choose three unit areas at a point, one normal to each of the three coordinate axes, and give the components of the force acting across each. Thus for the surfaces normal to the x , y , and z axes, we have the three force vectors, or nine quantities,

$$\begin{array}{ccc} X_x & Y_x & Z_x \\ X_y & Y_y & Z_y \\ X_z & Y_z & Z_z \end{array} \quad (1)$$

We see in Fig. 23 the significance of the three components X_x , Y_x , Z_x . This set of nine quantities forms the so-called stress tensor. The diagonal terms of the array, X_x , Y_y , Z_z , are called

the normal stresses or pressures, since the force components act normal to the surface, and the remaining terms are called shearing or tangential stresses. It is easily shown that the force across an arbitrary surface which has direction cosines l, m, n for its normal has an x component $lX_x + mX_y + nX_z$, with corresponding formulas for the other components.

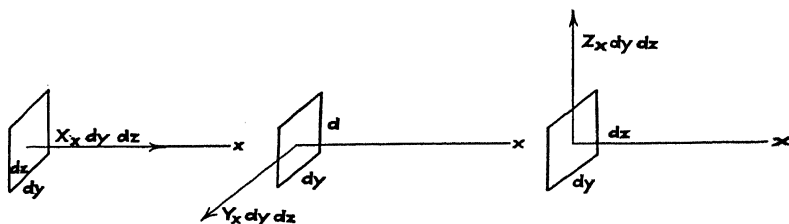


FIG. 23.—Components of force acting across $dydz$.

107. Examples of Stresses.—The simplest stress is probably the hydrostatic pressure. There the force acting across a square centimeter is always at right angles to the area, and its magnitude is by definition the pressure P . The force acts into the body, and hence is of negative sign. We thus have $X_x = Y_y = Z_z = -P$, all other components $= 0$. A second example is a tension, say in the x direction. Then the unit area perpendicular to x has a force T exerted across it, normal to the area, but there is no force exerted across faces perpendicular to y or z . In other words, $X_x = T$, all other components of the stress are zero. A third example is a shear. In Fig. 24 *a*, we have a cube of material, with equal and opposite tangential forces exerted across the faces normal to x , the forces acting in the y direction. Over the right face, the force exerted on the material is in the $-y$ direction, so that for this face we have $Y_x = -S$, a constant, and $X_x = Z_x = 0$. Over the opposite face, both force and direction of normal are reversed, so that the stress components are unchanged. But now we notice an important feature of shearing stress: the two forces we have mentioned exert a torque or couple on the cube, and if they were the only forces acting, it could not be in equilibrium. To get equilibrium, it proves necessary to have at the same time tangential forces exerted across the faces perpendicular to the y axis, as in Fig. 24 *b*. These forces are equal in magnitude to the other, so that the torques obviously balance, and we have $X_y = Y_x = -S$, all other components equal to

zero. This property, that $X_y = Y_x$, proves to be general: the stress tensor is symmetrical about its diagonal.

By making a proper rotation of axes, it is always possible to reduce a stress to diagonal form, in which no shearing stresses appear. Thus, in the case we have just considered, the problem is obviously symmetrical about the diagonal of the cube. In Fig. 24c, we take a surface element whose normal has direction cosines $l = -1/\sqrt{2}$, $m = 1/\sqrt{2}$, $n = 0$, giving a force exerted across it of components $-S/\sqrt{2}$, $S/\sqrt{2}$, 0, or a force of magnitude S normal to the surface. Similarly in Fig. 24d, we have a

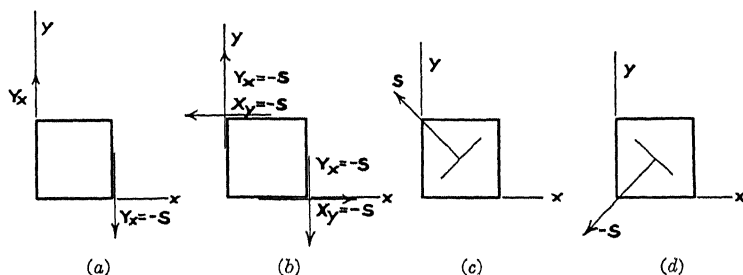


FIG. 24.—Diagram of shearing stress.

- (a) Shear over the faces perpendicular to the x axis.
 (b) Additional shear over faces perpendicular to the y axis, necessary to balance the turning moment of the shear indicated in (a).
 (c) and (d) Stress system of (b) referred to principal axes, tension in (c), pressure in (d).

surface at right angles, and find again a force normal to the surface, but now of magnitude $-S$. Thus, if we take as new axes the two 45-deg. diagonals in the xy plane, and the z axis, the stress consists of a tension S along one axis, negative tension (or pressure-like force) at right angles, and zero stress across the face normal to z . Axes of this sort, in which each face has a pure pressure- or tension-like force across it, and no shear, are called principal axes of stress.

108. The Equation of Motion.—Let us find the force on a small element of volume, having sides dx , dy , dz . Over the face at $x + dx$, there will be a force $X_x(x + dx)$, $Y_x(x + dx)$, $Z_x(x + dx)$ per unit area. Similarly exerted over the face at x there will be a force $-X_x(x)$, $-Y_x(x)$, $-Z_x(x)$. The x component of the resulting force is $X_x(x + dx) - X_x(x) = \frac{\partial X_x}{\partial x} dx$ per unit area,

or $\frac{\partial X_x}{\partial x} dx dy dz$ for the area $dy dz$. The y and z components are $\frac{\partial Y_x}{\partial x} dx dy dz$ and $\frac{\partial Z_x}{\partial x} dx dy dz$, respectively. In the same way we can find the three components of force exerted over each of the two other pairs of faces. Adding, we have for the total x component of force $\left(\frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z} \right) dx dy dz$. Thus, if v_x, v_y, v_z are the components of velocity of the solid at the point in question, the equations of motion, remembering that the mass of our small volume is $\rho dx dy dz$, are

$$\begin{aligned}\frac{\partial X_x}{\partial x} + \frac{\partial X_y}{\partial y} + \frac{\partial X_z}{\partial z} &= \rho \frac{dv_x}{dt} \\ \frac{\partial Y_x}{\partial x} + \frac{\partial Y_y}{\partial y} + \frac{\partial Y_z}{\partial z} &= \rho \frac{dv_y}{dt} \\ \frac{\partial Z_x}{\partial x} + \frac{\partial Z_y}{\partial y} + \frac{\partial Z_z}{\partial z} &= \rho \frac{dv_z}{dt}.\end{aligned}\tag{2}$$

These equations are evidently simply the generalization of those used previously with the string and membrane. Thus with the membrane let the z axis be normal to the plane of the membrane. We consider then only the third equation, giving velocity along z . The stress is a tension along the membrane, and if we cut the membrane with a surface perpendicular to x , we see that, if the membrane is inclined so that it makes an angle α with the x axis, there will be a component Z_x , a force in the direction to produce acceleration, equal to $T\alpha$. If then $\alpha = \partial u / \partial x$, where u is the displacement along z , the first term becomes $T(\partial^2 u / \partial x^2)$, as we found before. Similarly the second term is $T(\partial^2 u / \partial y^2)$, and the third is zero, yielding the equation of vibration which we have already used.

109. Transverse Waves.—Two sorts of waves are possible in an elastic solid: transverse waves, in which the displacement is at right angles to the direction of propagation of the waves, and longitudinal waves, as the sound waves in a gas, in which the displacement is in the direction of propagation. We consider first transverse waves. Rather than taking the general case, which involves rather complicated formulas, we assume that our wave is being propagated along the x axis, and that the displacement of the particles is in the y direction. We shall expect to get a wave equation involving only x derivatives, not y or z , and

having as solutions either progressive or standing waves. Let the displacement of a particle in the y direction be η ; since the wave is being propagated along the x axis, we assume that it has wave fronts normal to x , such that every point on a wave front has the same displacement, and this means that η is a function of x only. We may then consider a thin sheet or lamina, as that between x and $x + dx$ in Fig. 25. Let us suppose that the two points which in the unstrained medium were at x and $x + dx$, $y = 0$, are displaced to the points P and P' , at distances $\eta(x)$ and $\eta(x + dx)$, respectively, from the axis. Then evidently the lamina has been sheared, and we must find the relation between the shearing stress and the strain (that is, displacement) which

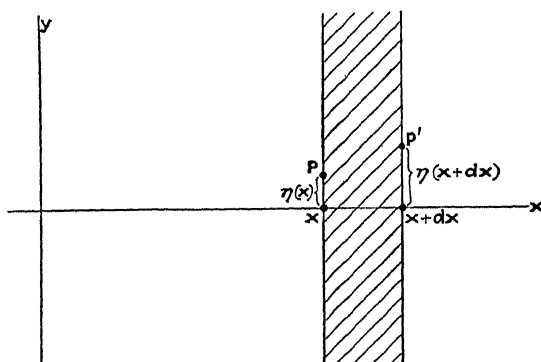


FIG. 25.—Shear in a transverse plane wave.

it has produced. The type of stress is evidently the sort described in Fig. 24. The material to the right of $x + dx$ exerts across unit cross section of the face a force in the y direction, equal to Y_x (or X_y). But now Hooke's law says that the actual deformation of the material, or the strain, is proportional to the stress acting. In this particular case, the deformation is a shearing one, and is opposed by the rigidity of the medium (which is the reason why a liquid, having no rigidity, cannot have transverse waves). The deformation is given in terms of the coefficient of rigidity μ as follows: the strain, measured by the tangent of the angle which the line PP' makes with the x axis, is equal to the shearing stress divided by μ . In other words, $Y_x = \mu \partial\eta/\partial x$. Substituting this relation between stress and strain in the equations of motion, we have at once

$$\frac{\partial}{\partial x} \left(\mu \frac{\partial \eta}{\partial x} \right) = \rho \frac{dv_y}{dt},$$

or, writing $v_y = \partial\eta/\partial t$,

$$\frac{\partial^2\eta}{\partial x^2} = \frac{\rho}{\mu} \frac{\partial^2\eta}{\partial t^2}, \quad (3)$$

the one-dimensional wave equation, representing transverse waves propagated with the velocity $\sqrt{\mu/\rho}$, or the square root of elastic modulus divided by density. Of course, we should have got the three-dimensional wave equation if we had considered propagation in an arbitrary direction.

110. Longitudinal Waves.—Here again we consider propagation along the x direction. In Fig. 26, let the displacement of a particle in the x direction be $\xi(x)$, a function of x only. Evidently the stress in this case is a pure tension, positive or negative, so

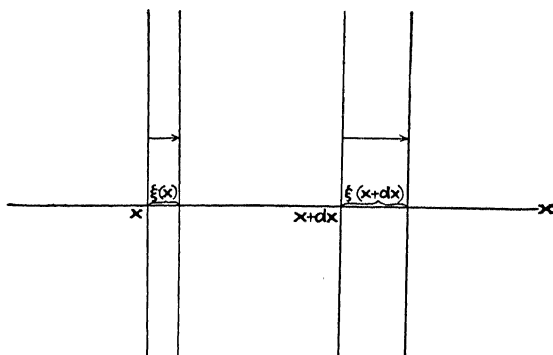


FIG. 26.—Compression and rarefaction in a longitudinal plane wave.

that the force across unit cross section is a pull in the x direction, equal to X_x . Hooke's law now states that the tension is proportional to the strain; and in particular, that it is proportional to the change in thickness of the lamina [which is evidently $\xi(x+dx) - \xi(x)$] divided by the thickness. The constant of proportionality in this case is not one of the simple elastic constants; it proves to be written $(\lambda + 2\mu)$, where λ is an elastic constant whose physical meaning is not easy to state. Perhaps as good an interpretation of λ as any is simply to define it from this particular sort of deformation. We now have $X_x = (\lambda + 2\mu) \frac{\partial\xi}{\partial x}$, all other components of stress = 0, so that from the equations of motion we at once have

$$\frac{\partial}{\partial x} \left[(\lambda + 2\mu) \frac{\partial \xi}{\partial x} \right] = \rho \frac{dv_x}{dt},$$

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{\rho}{\lambda + 2\mu} \frac{\partial^2 \xi}{\partial t^2}, \quad (4)$$

again a wave equation, representing a longitudinal wave traveling with velocity $\sqrt{(\lambda + 2\mu)/\rho}$, different from the velocity of the transverse wave.

111. General Wave Propagation.—In the two preceding sections, we have derived two very specialized waves which can be propagated in an elastic solid, plane longitudinal and transverse waves traveling along the x axis. Of course, much more complicated waves are possible, and if we were discussing the problem completely, we should set up the three-dimensional wave equation, of the form $\nabla^2 u = \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2}$, and derive general

wave solutions. We should have separate equations for the longitudinal and transverse waves, generalizations of Eqs.

(3) and (4). As we shall learn later when discussing optical problems, such a wave equation has as solutions not merely plane waves traveling in all arbitrary directions, but also spherical waves diverging from

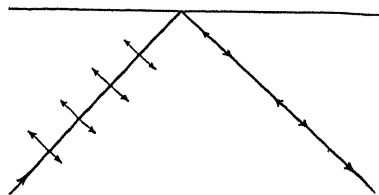


FIG. 27.—Incident transverse wave, with longitudinal reflected wave.

point sources, and many more complicated types of waves. All these are possible in an elastic solid. In our discussion of the plane waves, we separated the longitudinal and transverse waves entirely, allowing one type to exist without the other, but unfortunately in general this cannot be done. For instance, when a wave of one type is reflected from a surface, then unless the reflection is at normal incidence, longitudinal motion will generally be partly converted into transverse, and *vice versa*. In Fig. 27, we show diagrammatically how this could be, the transverse motion in the incident wave evidently being in such a direction as to be partly transformed into longitudinal motion in the reflected wave. For this reason, the complete treatment of the vibrations of an elastic solid is a very complicated problem. An example is found in geophysical problems, where one is interested in the propagation of earthquake waves through the earth. This case is made even more difficult by the fact

that the elastic properties of the earth change as a function of depth, so that one must use solutions of the form we have discussed in Chap. XIV, in connection with strings whose properties depend on position.

There is one application of the theory of the waves in an elastic solid which has at least historical interest. When it was discovered that light was a transverse wave motion, it was attempted to identify these waves with the transverse vibrations of an elastic solid, the ether. The general properties, and even some of the details, as the quantitative laws giving the fraction of light reflected and transmitted at a boundary, were correctly worked out, the reflection being treated by analogy with our discussion of reflection of waves in strings at a point of discontinuity of density, in Chap. XIV. But the difficulty, which could not be overcome, was that of eliminating the longitudinal waves, which certainly do not occur in optics, but which were inherent in the elastic solid theory. This difficulty does not occur in the present electromagnetic theory, where only transverse waves are allowed by the fundamental differential equations. This lack of longitudinal waves makes the problem of optical wave motion on the whole simpler than that of elastic waves.

112. Strains and Hooke's Law.—In discussing transverse and longitudinal elastic waves, we had to introduce certain elastic constants, measuring the ratio between stress components, and certain quantities measuring the strain or deformation of the substance. The fact that these strains were proportional to the stresses is Hooke's law, the fundamental law of elasticity, holding for sufficiently small strains. It is now worth while to state the general relation between stress and strain, though we shall not go through the proof.

To begin with, we imagine the body unstrained. Then in the process of deformation, we imagine that the particle originally at x, y, z has been displaced to a point $x + \xi, y + \eta, z + \zeta$. The three quantities ξ, η, ζ are functions of x, y, z , and are the three components of a vector. We meet, in other words, a vector (which we may call the displacement), which is a function of position. Such a vector field reminds us of a force field, as a gravitational- or electric-force field, where the force vector on unit mass or charge, respectively, is a function of position. We shall meet such vector fields often in the future. Now, the displacement is not the same thing as the strain; the body might

be displaced bodily, without involving any stress or strain at all. It is only when the displacement of one side of a small element of volume is different from the other, so that the element is distorted in size or shape, that we have a strain. In other words, the essential quantities in determining the strain are the derivatives of ξ , η , ζ with respect to x , y , z . We have already seen two examples: with the shear in the transverse wave, the strain was $\partial\eta/\partial x$, and in the compressional wave the strain was $\partial\xi/\partial x$. In the two cases mentioned, the stress was proportional to the corresponding partial derivative, and Hooke's law means that this is true in general, in the form that the components of stress are linear functions of the partial derivatives of the components of displacement. There are nine components of stress, of which six are independent (remembering that $X_y = Y_x$, etc.), and similarly there are nine partial derivatives of displacement, of which it can be proved that six again are independent. This would mean six linear equations, with thirty-six coefficients, which would act as elastic constants. In the most general type of substance, a completely anisotropic crystal, it can be shown that twenty-one of these really are independent, giving a tremendous number of elastic constants. With isotropic substances showing no crystalline structure, however, most of these constants are either zero or can be written in terms of each other, and there are only two independent constants, the λ and μ which we have already met; all other elastic constants, as Young's modulus and the compressibility, can be written in terms of them. Using these constants, the relations between stress and strain prove to have the following form:

$$\begin{aligned} X_x &= (2\mu + \lambda) \frac{\partial\xi}{\partial x} + \lambda \frac{\partial\eta}{\partial y} + \lambda \frac{\partial\zeta}{\partial z} & X_y &= \mu \left(\frac{\partial\xi}{\partial y} + \frac{\partial\eta}{\partial x} \right) \\ Y_y &= \lambda \frac{\partial\xi}{\partial x} + (2\mu + \lambda) \frac{\partial\eta}{\partial y} + \lambda \frac{\partial\zeta}{\partial z} & Y_z &= \mu \left(\frac{\partial\eta}{\partial z} + \frac{\partial\zeta}{\partial y} \right) \\ Z_z &= \lambda \frac{\partial\xi}{\partial x} + \lambda \frac{\partial\eta}{\partial y} + (2\mu + \lambda) \frac{\partial\zeta}{\partial z} & Z_x &= \mu \left(\frac{\partial\zeta}{\partial x} + \frac{\partial\xi}{\partial z} \right). \end{aligned} \quad (5)$$

In the cases we have taken up already, we have seen two illustrations of these equations: with transverse waves, $\partial\eta/\partial x$ was the only partial derivative different from zero, and we had $X_y = \mu \partial\eta/\partial x$; with the longitudinal wave, $\partial\xi/\partial x$ was the only term different from zero, and as we see this gives $X_x = (2\mu + \lambda) \partial\xi/\partial x$, as we had before, but also $Y_y = Z_z = \lambda \partial\xi/\partial x$. These latter

stress components, however, since they do not depend on y or z , do not contribute to the equations of motion, as we see by referring back to these.

113. Young's Modulus.—To illustrate the use of the equations connecting stress and strain, we shall discuss the stretching of a wire. Let the wire be stretched along the x axis, and let the stress be a pure tension T , so that $X_x = T$, and all other stress components are zero. The x , y , z axes are principal axes for this stress, and it can be shown that the strain has principal axes, too, parallel to those of stress, so that the last three equations, for X_y , etc., do not enter. We are left, then, with the three equations

$$\begin{aligned} T &= (2\mu + \lambda) \frac{\partial \xi}{\partial x} + \lambda \frac{\partial \eta}{\partial y} + \lambda \frac{\partial \zeta}{\partial z} \\ 0 &= \lambda \frac{\partial \xi}{\partial x} + (2\mu + \lambda) \frac{\partial \eta}{\partial y} + \lambda \frac{\partial \zeta}{\partial z} \\ 0 &= \lambda \frac{\partial \xi}{\partial x} + \lambda \frac{\partial \eta}{\partial y} + (2\mu + \lambda) \frac{\partial \zeta}{\partial z} \end{aligned}$$

Subtracting the third from the second, we have $\partial \eta / \partial y = \partial \zeta / \partial z$. Using this relation, either the second or third gives $\partial \eta / \partial y =$

$$-\sigma \left(\frac{\partial \xi}{\partial x} \right), \text{ where } \sigma = \frac{\lambda}{2(\lambda + \mu)}, \text{ and is called Poisson's ratio.}$$

Since λ and μ are always positive, it is obvious that Poisson's ratio is never greater than $1/2$. We have found, then, that as the wire is stretched (positive $\partial \xi / \partial x$), it contracts sidewise, (negative $\partial \eta / \partial y$ and $\partial \zeta / \partial z$) and the ratio of sidewise contraction per unit width, to lengthwise stretch per unit length, is given by Poisson's ratio. Actual materials have Poisson's ratio of the order of magnitude of $1/3$. Now we put this expression back in the first equation, obtaining $T = (2\mu + \lambda - 2\lambda\sigma) \partial \xi / \partial x$. The elastic modulus $(2\mu + \lambda - 2\lambda\sigma)$, giving the tension, or force per unit area, divided by the elongation per unit length, is called Young's modulus, and is denoted by E . In the problems we find other ways of writing the relations between Young's modulus, Poisson's ratio, and the other elastic constants.

It is worth noticing that Young's modulus was not the elastic constant which entered into the velocity of compressional waves. If we had longitudinal waves traveling down a wire, the wire would contract laterally at those points where it was under tension, expand when it was under compression, as given by

Poisson's ratio, and for such a wave the velocity would be determined from Young's modulus. But in our extended medium, we did not allow the possibility of the lateral motion connected with such a contraction and expansion, since in a medium of large dimensions compared with the wave length this would amount to a very large transverse motion. We assumed instead that the motion was purely longitudinal, and found that we had to assume the existence of other lateral stresses, tensions Y_y and Z_z , to counteract the tendency to expansion and contraction. These stresses changed the conditions of the problem, and in particular the elastic modulus concerned in the velocity of propagation of the wave.

Problems

1. In Fig. 28, let the normal to the inclined face of the prism have direction cosines l, m, n . Compute the total forces exerted by an arbitrary stress

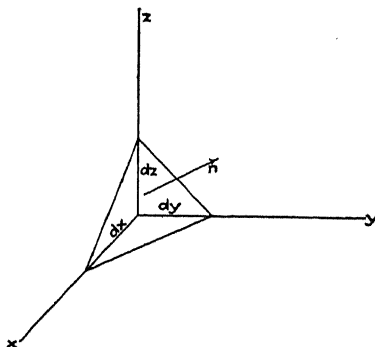


FIG. 28.—Prism for computing force exerted by stresses across a face with arbitrary normal n .

on the prism, and prove that the net force is zero, and the prism is in equilibrium, only if the force per unit area over the face perpendicular to n has x component $lX_x + mX_y + nX_z$, etc.

2. Rotate coordinates to reduce an arbitrary stress to principal axes. Carry through the problem of the pure shear, discussed in Fig. 24, as an illustration of the general method.

3. Prove that in terms of Young's modulus and Poisson's ratio we have

$$\lambda = \frac{E\sigma}{(1+\sigma)(1-2\sigma)}, \quad 2\mu = \frac{E}{1+\sigma}.$$

4. Assume a body is under pure hydrostatic pressure P . Show that the distortion is a decrease of all dimensions by a fixed fraction. Show that the fractional change in volume is $\partial\xi/\partial x + \partial\eta/\partial y + \partial\zeta/\partial z$. Using this, show that the compressibility κ of a solid under hydrostatic pressure, which

by definition is the fractional decrease of volume divided by the pressure, equals $3(1 - 2\sigma)/E$.

5. Show that the velocity of a longitudinal wave in a fluid, for which μ is zero, is $1/\sqrt{\kappa\rho}$, where κ is the compressibility.

6. A rectangular beam held at one end is bent into an arc of a circle, the radius of curvature of its central section being R . Find the stress distribution throughout the beam, showing that the beam will be kept in equilibrium by a torque or couple of the sort indicated. Show that for a given torque the curvature of the beam is inversely proportional to ab^3E , where E is Young's modulus (see Fig. 29).

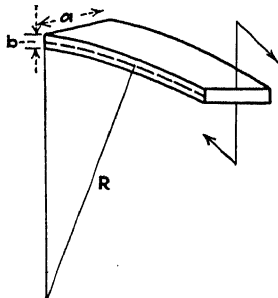


FIG. 29.—Bent beam.

7. A circular cylinder of height h rests on a flat surface under the action of gravity. Take a coordinate system with the xy plane in the top base of the cylinder and the positive z axis pointing downward. Show that the only component of stress different from zero is $Z_z = -\rho gz$, if ρ is the density of the cylinder. Using Hooke's law show that the strains are $\partial\xi/\partial x = \partial\eta/\partial y = (\sigma/E)\rho gz$, and $\partial\zeta/\partial z = -(1/E)\rho gz$, and find the other partial derivatives. Integrate these expressions to find the components of the displacement of any point of the medium, remembering that the strains are partial derivatives. Show that a horizontal plane section of the cylinder becomes a paraboloid of rotation due to the deformation. Show that the radius of the cylinder increases from top to bottom when it is thus deformed.

8. A spherical shell of inner radius R_1 , outer radius R_2 , contains a fluid of pressure P_1 , and is immersed in a second fluid of lower pressure P_2 . It can be shown that the displacements of points on account of the pressure are given by $\xi = x(A + B/r^3)$, $\eta = y(A + B/r^3)$, $\zeta = z(A + B/r^3)$. Verify these values by computing the stresses at any point, substituting in the equations of motion, and showing that they result in equilibrium. Show further that the force across an area normal to the radius is itself normal to the surface, so that the stress within the sphere can be balanced by hydrostatic pressures within and without.

9. In the shell of Prob. 8, determine A and B so that the pressure will have the proper values at R_1 and R_2 . Discuss the stress within the shell, showing that the principal axes at any point are along the radius and two arbitrary directions at right angles, and find the tension or pressure along the directions at right angles, discussing the final result physically, with special reference to possible breaking of the shell under excessive pressure inside.

CHAPTER XVII

FLOW OF FLUIDS

In the last chapter we discussed the equation of motion of an elastic body where there was no mass motion or flow. Now we pass to hydrodynamics and the flow of fluids. Much of what we say, however, applies to flow in general—such as heat flow, which we shall take up in the next chapter—and even to such a different subject as electrostatics. The feature in common in all these problems is the existence of a vector field. By that we mean a vector defined at each point of space. We have already met such a field in our general discussion of forces and potentials in Chap. VI, for the force is defined at every point of space and forms a vector field. In the present case the vector is the velocity of the flowing fluid, or the closely related flux density. With heat flow it is again a flux density for the flowing heat, and for electricity the electric field. All these problems, though so different physically, are thus mathematically similar and can be treated by the same analytical methods.

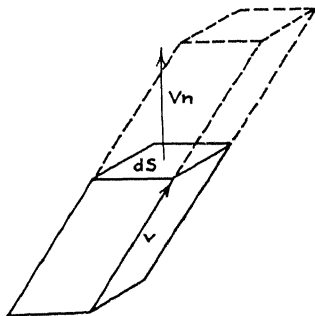


FIG. 30.—Flux through an area dS .

114. Velocity, Flux Density, and Lines of Flow.—At every point of a flowing medium, we can define the velocity, a vector (the time rate of change of the displacement, which we used in the last chapter, and to which we assigned components ξ, η, ζ). Also we can give the density ρ , and both ρ and v are in general functions of position (x, y, z) and of time. We may now ask, How much material will flow across any area per second? This total flow across a surface is called a flux. In Fig. 30, we consider an infinitesimal surface element dS . With dS as a base we erect a prism, the slant height being the velocity v , which in general is not normal to dS . Evidently the material in the prism will just be that which crosses dS in one second, since in

this time it will move a distance v , and fill the dotted prism. But this is ρ (the density) times the volume of the prism (the base dS times the altitude v_n , where n is the normal to the surface), or $\rho v_n dS$. The quantity ρv is called the flux density, and we may denote it by f . Then for a finite area, the total flux will be the sum of the contributions from all the surface elements, or a surface integral $\iint f_n dS = \iint \rho v_n dS$. In some kinds of flow, such as heat flow, there is an analogue to the flux, but not to the density and velocity separately, so that one regards the flux density as being the more fundamental vector field.

We can draw lines through the medium, tangent at every point to the direction of flow at that point. These are called lines of flow. Similarly we can set up tubes of flow, the elements of their surfaces being lines of flow. We can imagine the substance to flow through these tubes, as water flows through a pipe, never passing outside, since the velocity is always tangential to the surface of the tube. In hydrodynamics these lines of flow are called streamlines, and the sort of flow in which they are independent of time is called streamline flow.

115. The Equation of Continuity.—Consider a fixed volume in a flowing fluid. The amount of fluid in the volume is $\iiint \rho dv$, and this can change in two ways. First, liquid can flow into the volume over the surfaces. Secondly, it may be possible for liquid to be produced within the volume without having flowed in. For instance, in a swimming pool, for all practical purposes we may consider the opening of the inlet pipe as a region where fluid is appearing, and the outlet as a place where it is disappearing. Such regions are called sources and sinks, respectively. Then we have

$$\iint \iint \frac{\partial \rho}{\partial t} dv = \text{rate of inflow over the surface} + \\ \text{rate of production inside.}$$

Now we have just seen that the rate of flow over any surface, or flux, is $\iint f_n dS$. This represents outflow if n is the outer normal to a closed surface, so that we must change sign to get inflow. If in addition we assume that the rate of production of material per unit volume is P , we have

$$\iint \iint \frac{\partial \rho}{\partial t} dv = - \iint f_n dS + \iint \iint P dv,$$

the volume integrals being over the whole region we are considering, the surface integral over the surface enclosing this volume.

If we now apply our equation to an infinitesimal volume in the form of a rectangular parallelepiped, bounded by $x, x + dx, y, y + dy, z, z + dz$, we can put the equation in a form not involving integrals. The flow to the right (into the volume) over the face x is $f_x(x)dydz$. The net flow over that at $x + dx$ is

$$f_x(x + dx)dydz = f_x(x)dydz + \frac{\partial}{\partial x}f_x(x)dydz \dots$$

Thus the total inflow over the faces is $-\frac{\partial}{\partial x}(f_x)dx dydz$. Adding similar contributions from the other faces we have for the total inflow

$$-\iint f_n dS = -\left(\frac{\partial}{\partial x}f_x + \frac{\partial}{\partial y}f_y + \frac{\partial}{\partial z}f_z\right)dx dydz = -(\nabla \cdot f)dv = -\text{div } f dv,$$

where the divergence is a vector operator discussed in Chap. VI. Hence

$$\frac{\partial \rho}{\partial t} = -\text{div } f + P. \quad (1)$$

This is often called the equation of continuity. We may note several special cases. If there is no production of fluid in dv , it becomes

$$\frac{\partial \rho}{\partial t} + \text{div } f = 0,$$

or using $f = \rho v$,

$$\frac{\partial \rho}{\partial t} + \text{div } (\rho v) = 0. \quad (2)$$

Again, in a steady state, where density is independent of time,

$$\text{div } f = P. \quad (3)$$

This equation shows the physical meaning of the divergence of a vector: it measures the rate of production of the flowing substance, per unit volume. Finally, if no substance is being produced at the point in question, and density is independent of time, $\text{div } f = 0$, and we have a divergenceless flow.

116. Gauss's Theorem.—We have proved that the amount of substance flowing out of a small volume $dx dy dz = dv$ per second equals $\text{div } f dv$ in steady flow. Suppose now that we have a large

volume and that we wish to find the total amount flowing out of it per second. This is simply the sum of the amounts flowing from each element. Thus it is a volume integral, $\iiint \text{div } f \, dv$. On the other hand, the material all flows through the surface, so that the rate of outflow is $\iint f_n dS$. These two expressions must be equal:

$$\iiint \text{div } f \, dv = \iint f_n dS. \quad (4)$$

This is Gauss's theorem, and it holds for any vector f which is a function of position.

117. Lines of Flow to Measure Rate of Flow.—Let us set up a definite number of lines of flow, so that the number crossing a unit area perpendicular to the flow is numerically equal to the magnitude of the flux density. We could surely do this, but we might have the necessity of sometimes letting lines start or stop, to keep the right number. We can prove, however, that with a divergenceless flow this would not be necessary. The lines start or stop only at places where the divergence is different from zero: that is, they start at sources, stop at sinks. For an elementary proof, let us take a short section of a tube of flow, bounded by two surfaces normal to the flow. Let one of them have an area A_1 , the other A_2 , and let the magnitude of the flux over the one face be f_1 , over the other f_2 . Then the total current in over one face is $f_1 A_1$, and out over the other is $f_2 A_2$. If the flow is divergenceless, these are equal. But the number of lines per unit area on the first is f_1 , so that the number cutting the one end of the tube is $f_1 A_1$, and the number emerging at the other end is $f_2 A_2$. Since these are equal, no lines are lost or start within. In other words, in a divergenceless flow, lines never start or stop except at sources or sinks. For a more general proof we note that the number of lines crossing a surface element dS , by definition, is $f_n dS$. Then the number emerging from a closed surface, and which therefore have started within the surface, is $\iint f_n dS$. But by Gauss's theorem this is $\iiint \text{div } f \, dv$, and is zero if the flow is divergenceless.

118. Irrotational Flow and the Velocity Potential.—In Chap. VI we studied vector fields like our flux vector; we were interested then in forces. We saw that under certain conditions, a force could be written as a gradient of a potential function. The condition was that the work done in taking a particle around any closed path should be zero, or that the field should be conserva-

tive: $\int \mathbf{F} \cdot d\mathbf{s} = 0$ around any contour. We had another way of stating the condition: it was $\text{curl } \mathbf{F} = 0$ everywhere. In a similar way, if the curl of our velocity vector is zero, we can introduce a potential function here. It is now to be regarded as a purely mathematical device, used simply by analogy with our previous cases, and having nothing to do with potential energy. A flow whose curl is everywhere zero is called an irrotational flow. It is easy to prove that in a whirlpool the curl is different from

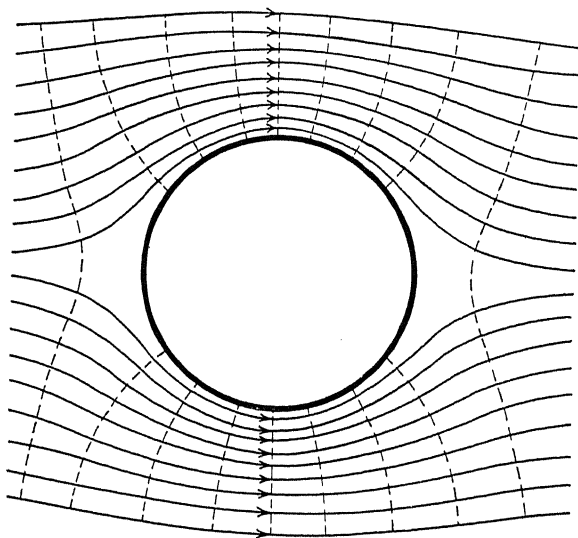


FIG. 31.—Lines of flow and equipotentials for flow about a cylinder. Full lines indicate lines of flow, dotted lines equipotentials. In a corresponding electrical problem with charges distributed within the cylinder, and placed in a uniform external electric field, the dotted lines would be lines of force, full lines equipotentials.

zero (see for instance Prob. 4, Chap. VI), a nonvanishing curl indicating in fact exactly a whirlpool. Now, physically, we are acquainted with two sorts of fluid flow: streamline flow and turbulent flow. In the latter, eddies or whirlpools form, and the curl of the velocity is not zero. But in the former, there are no eddies, the curl of the velocity is zero, and the flow is irrotational. In a streamline flow, then, we can introduce a potential function, called the velocity potential ϕ , defined by $v = -\text{grad } \phi$. The velocity potential, of course, is not a potential energy; its analogy with potential energies is mathematical rather than physical. Nevertheless, we can draw surfaces of constant velocity potential,

or equipotentials, and the lines of flow will cut the equipotentials at right angles. Using the equation of continuity, and assuming that ρ is constant, we have as the general equation for the velocity potential

$$\operatorname{div}(\rho \mathbf{v}) = -\rho \operatorname{div} \operatorname{grad} \phi = -\rho \nabla^2 \phi = -\frac{\partial \rho}{\partial t} + P. \quad (4)$$

reducing to Laplace's equation $\nabla^2 \phi = 0$ for a steady state where there are no sources or sinks.

The introduction of a velocity potential satisfying Laplace's equation makes it possible in many cases to solve hydrodynamic problems by analogy with similar problems in other branches of physics, as electrostatics. In Chap. XIX we shall find that the electrostatic potential satisfies Laplace's equation, the lines of force being normal to the equipotentials, so that any set of electrostatic equipotentials can be used for a suitable hydrodynamic problem. For instance, in Fig. 31, we show the lines of flow and equipotentials for flow of a liquid about a cylinder. The same lines, however, represent lines of force resulting from a certain distribution of charges in the center of the sphere, superposed on a uniform electric field.

119. Euler's Equations of Motion for Ideal Fluids.—The equation of continuity serves to determine the velocity of flow of a liquid, but does not determine the pressures, or make any connection with forces. It is essentially a kinematical rather than a dynamical law. It is one of two fundamental equations governing fluid motion. The other is essentially the Newtonian law, force equals mass times acceleration. For a continuous medium, we have already seen how this is to be formulated in the preceding chapter, where we wrote the force on an element of volume in terms of the stresses. As was mentioned in the last chapter, an ideal fluid is characterized by the fact that it supports no shear and hence $\mu = 0$. For this case the six stress components reduce to one, namely $X_x = Y_y = Z_z = -p$ and $X_y = Y_x = Y_z = Z_y = Z_x = 0$, if p denotes the pressure in the fluid. Furthermore, if there is flow of the fluid one must consider the velocity of each particle as a function of x, y, z , and t , and hence

$$\frac{dv_x}{dt} = \frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} + v_z \frac{\partial v_x}{\partial z}$$

and two similar expressions for v_y and v_z . Written in vector form with the help of our symbolic vector $\nabla = \operatorname{grad}$

$$\frac{dv_x}{dt} = \frac{\partial v_x}{\partial t} + (v \cdot \nabla)v_x = \frac{\partial v_x}{\partial t} + (v \cdot \text{grad})v_x,$$

i.e., we form the scalar product of v and ∇ and then operate on v_x . Our general equations of motion become in this case:

$$\begin{aligned}\rho X - \frac{\partial p}{\partial x} &= \rho \left\{ \frac{\partial v_x}{\partial t} + (v \cdot \text{grad})v_x \right\} \\ \rho Y - \frac{\partial p}{\partial y} &= \rho \left\{ \frac{\partial v_y}{\partial t} + (v \cdot \text{grad})v_y \right\} \\ \rho Z - \frac{\partial p}{\partial z} &= \rho \left\{ \frac{\partial v_z}{\partial t} + (v \cdot \text{grad})v_z \right\}.\end{aligned}$$

where X , Y , Z represent the body force (as gravitation) per unit mass, which we neglected in the last chapter. Combined into one vector equation this gives

$$F - \frac{1}{\rho} \text{grad } p = \frac{\partial v}{\partial t} + (v \cdot \text{grad})v, \quad (5)$$

where F is the body force. These are the Euler equations of hydrodynamics. In them ρ (the density) is considered a known function of the pressure as given by the equation of state of the substance. We then have p , v_x , v_y , v_z as functions of x , y , z , and t . The three equations above and the continuity equation provide the necessary four equations to give a unique solution. For the case of hydrostatic equilibrium, these equations reduce to the form $F = (1/\rho) \text{grad } p$, from which such familiar things as Archimedes' principle immediately follow.

120. Irrotational Flow and Bernoulli's Equation.—If there is irrotational flow, and the velocity is derived from a velocity potential, Euler's equations take a particularly simple form. If $v = -\text{grad } \phi$, then we have

$$\begin{aligned}(v \cdot \text{grad})v_x &= -\left(v \cdot \text{grad} \frac{\partial \phi}{\partial x}\right) \\ &= \frac{\partial \phi}{\partial x} \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial \phi}{\partial y} \frac{\partial^2 \phi}{\partial x \partial y} + \frac{\partial \phi}{\partial z} \frac{\partial^2 \phi}{\partial x \partial z} \\ &= \frac{1}{2} \frac{\partial}{\partial x} \left[\left(\frac{\partial \phi}{\partial x} \right)^2 + \left(\frac{\partial \phi}{\partial y} \right)^2 + \left(\frac{\partial \phi}{\partial z} \right)^2 \right],\end{aligned}$$

so that

$$(v \cdot \text{grad})v = \text{grad} \left(\frac{v^2}{2} \right).$$

in the special case where $\text{curl } v = 0$. Further, we introduce a quantity Π , defined by the equation $\Pi = \int_0^p \frac{dp}{\rho(p)}$, whose gradient is

$$\text{grad } \Pi = \frac{d\Pi}{dp} \text{ grad } p = \frac{1}{\rho} \text{ grad } p.$$

Euler's equation for the steady state, where v is independent of time, then becomes

$$F = \text{grad} \left(\Pi + \frac{v^2}{2} \right).$$

As a result of this equation, we see that for irrotational flow to occur, F must be the gradient of a certain quantity, or F must be a conservative force, derivable from a potential. We may then set $F = -\text{grad } V$, and Euler's equation becomes

$$\text{grad} \left(V + \Pi + \frac{v^2}{2} \right) = 0,$$

or, integrated,

$$V + \Pi + \frac{v^2}{2} = \text{constant}.$$

This is Bernoulli's equation. For the special case of an incompressible fluid, ρ is independent of p , so that Π is equal to $\frac{p}{\rho}$. In that case the equation may be written

$$\rho V + p + \frac{1}{2} \rho v^2 = \text{constant}.$$

Bernoulli's equation is essentially an energy integral, the term ρV representing the potential energy per unit volume, p the contribution to the energy resulting from the pressure, and $\frac{1}{2} \rho v^2$ the kinetic energy per unit volume. As we have stated, Bernoulli's equation, supplemented for a compressible fluid by the relation giving density as function of pressure, determines the pressure at each point of space, when the velocity and external potential are known. For instance, if there is no external force field ($V = 0$), we see that the pressure decreases at points where the velocity is high, which means at points where the tubes of flow narrow down.

121. Viscous Fluids.—In Sec. 119 we mentioned the fact that ideal fluids support no shearing stresses. This, however, is not true of viscous fluids. Imagine a viscous liquid flowing hori-

zonally, the lower layers dragging along the bottom, and the velocity increasing with height, so that $v_x = v_x(y)$, other components of v are zero, if the xz plane is horizontal, y is vertical. Then if we imagine a horizontal element of area in the liquid at a certain height, the material above the element of area will pull tangentially on the material below it on account of viscosity, thus exerting a shearing stress. Experimentally, this stress, which is X_y , is proportional to the rate of increase of horizontal component of velocity with height: if k is the coefficient of viscosity, $X_y = k \frac{\partial v_x}{\partial y}$. This is a special case of the general laws

governing stresses in a viscous medium, connecting the stresses with the rates of change of the velocity components with position.

In the last chapter we have given the general form of Hooke's law, the law giving stresses in an elastic medium in terms of the strains. By analogy we can set up the relations for a viscous fluid, but now the stresses are proportional, not to the strain components themselves, but to their time derivatives. By comparison with Eq. (5), Chap. XVI, we see that k takes the place of the shear modulus, and that the component of strain $\partial \xi / \partial y + \partial \eta / \partial x$ must be replaced by its time derivative, $\partial v_x / \partial y + \partial v_y / \partial x = \partial v_x / \partial y$ in our special case, since $v_y = 0$. This tells us how in general we are to change Hooke's law for the case of viscous incompressible fluids. We place $\partial v_x / \partial x + \partial v_y / \partial y + \partial v_z / \partial z = \text{div } v = 0$, corresponding to $\partial \xi / \partial x + \partial \eta / \partial y + \partial \zeta / \partial z = 0$ in the strains, replace μ by k and insert the time derivatives of the strain components. Thus we have the following relations between the stress and strain components for liquids:

$$\begin{aligned} X_x &= -p + 2k \frac{\partial v_x}{\partial x}; & X_y &= k \left(\frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x} \right) \\ Y_y &= -p + 2k \frac{\partial v_y}{\partial y}; & Y_z &= k \left(\frac{\partial v_y}{\partial z} + \frac{\partial v_z}{\partial y} \right) \\ Z_z &= -p + 2k \frac{\partial v_z}{\partial z}; & Z_x &= k \left(\frac{\partial v_z}{\partial x} + \frac{\partial v_x}{\partial z} \right) \end{aligned} \quad (6)$$

where we have included the ordinary pressure of the liquid in addition to the viscous stresses. Inserting the values of the stress components in the equations of motion (2) of the previous chapter and remembering that for an incompressible fluid we have the continuity equation $\text{div } v = \partial v_x / \partial x + \partial v_y / \partial y +$

$\partial v_z / \partial z = 0$, there follow the general equations of motion for viscous liquids:

$$\begin{aligned}\rho X - \frac{\partial p}{\partial x} + k \nabla^2 v_x &= \rho \frac{dv_x}{dt} \\ \rho Y - \frac{\partial p}{\partial y} + k \nabla^2 v_y &= \rho \frac{dv_y}{dt} \\ \rho Z - \frac{\partial p}{\partial z} + k \nabla^2 v_z &= \rho \frac{dv_z}{dt}\end{aligned}\quad (7)$$

or in vector form: $\rho F - \text{grad } p + k \nabla^2 v = \rho \frac{dv}{dt}$, differing from Eq. (5) by the term $k \nabla^2 v$.

122. Poiseuille's Law.—Suppose we have an incompressible liquid flowing in a steady state in a horizontal cylinder of radius R parallel to the long axis of the cylinder (x axis). We have $v_y = v_z = 0$ and since there are no body forces $X = Y = Z = 0$. The equation of continuity becomes $\partial v_x / \partial x = 0$ so that v_x is a function of y and z alone. Then $dv_x / dt = v_x \partial v_x / \partial x + v_y \partial v_x / \partial y + v_z \partial v_x / \partial z = 0$. Furthermore, if we take the divergence of the fundamental equations of motion, we have:

$$\rho \text{ div } F - \text{div grad } p + k \nabla^2 (\text{div } v) = \rho \frac{d}{dt} (\text{div } v)$$

Now by the equation of continuity $\text{div } v = 0$, and in our case of no external forces this reduces to

$$\text{div grad } p = \nabla^2 p = 0.$$

In our problem $\partial p / \partial y = \partial p / \partial z = 0$, so that $d^2 p / dx^2 = 0$. The pressure is thus a linear function of x , so that we have a constant pressure gradient in the tube. Of the three equations, only the first is left:

$$\frac{dp}{dx} = k \left(\frac{\partial^2 v_x}{\partial y^2} + \frac{\partial^2 v_x}{\partial z^2} \right)$$

and since dp/dx is constant $= a$, and we have cylindrical symmetry, this reduces to

$$\frac{1}{r} \frac{d}{dr} \left(r \frac{dv_x}{dr} \right) = \frac{a}{k}$$

where r is the distance from the axis of the cylinder. Integrated, this yields $v_x = \frac{a}{4k} r^2 + b \ln r + c$, and since v_x is finite for

$r = 0$, $b = 0$. If the liquid clings to the walls of the cylinder, $v_x = 0$ when $r = R$, so that we find

$$v_x = \frac{a}{4k}(r^2 - R^2). \quad (8)$$

Thus the liquid flows in cylindrical tubes of constant velocity. This type of motion is called "laminar" motion. The velocity varies parabolically across a diameter of the cylinder.

The amount of liquid flowing per second through a cylindrical ring of thickness dr , radius r , is

$$dQ = 2\pi r v_x dr$$

so that the total discharge rate of such a cylinder is

$$Q = 2\pi \int_0^R r v_x dr = -\frac{\pi a R^4}{8k} = \frac{\pi R^4}{8kL} (p_2 - p_1) \quad (9)$$

where we have placed the constant pressure gradient $a = -\frac{p_2 - p_1}{L}$. This law, known as Poiseuille's law, furnishes a very nice experimental method of determining the coefficient of viscosity of liquids.

Problems

1. Liquid is confined between two parallel plates, so that it flows in two dimensions. At a certain point, a pipe discharges liquid at a constant rate into the region. Find the velocity potential, and velocity, as a function of position. Show by direct calculation that the flow outward over any circle about the source is the same.

2. A shallow tray containing fluid has a source at one point, an equal sink at another, so that liquid flows in two dimensions from source to sink. Find the equation of the equipotentials and the lines of flow, prove they are circles and plot them. (Suggestion: since the equations are linear, the potential or flux due to two sources is the sum of the solution for the separate sources.)

3. Prove that $\frac{\partial}{\partial x}(1/r)$ is a solution of Laplace's equation. Investigate the lines of flow connected with this as a potential. Draw the lines, in the xy plane. What sort of physical situation would be described by this case?

4. Consider an ideal fluid at rest. It is subjected to an impulsive pressure $(p) = \int_0^t \tau dt$, where τ indicates the interval of time during which the pressure is applied. If no body forces act on the fluid, prove by integrating Euler's equations, that the impulsive pressure divided by the density of the fluid equals the velocity potential of the ensuing motion. This is the physical significance of a velocity potential.

5. Show for a liquid in equilibrium under the action of gravity that the pressure varies linearly with the depth below the surface. Calculate the total force exerted on the surface of a submerged body by the liquid and show that the resultant force is directed upwards and is given in magnitude by Archimedes' principle. [Hint: If a vector has only one component different from zero, *e.g.*, A_x , then Gauss's theorem becomes

$$\int \frac{\partial A_x}{\partial x} dV = \int A_x \cos(n, x) dS.]$$

6. The free surface of a liquid is one of constant pressure. If an incompressible fluid is placed in a cylindrical vessel and the whole rotated with constant angular velocity ω , show that the free surface becomes a paraboloid of revolution. (Hint: Introduce a fictitious potential energy to take care of centrifugal force and use the hydrostatic equations.)

7. A gas maintained at constant pressure p , flows steadily out of a small hole into the atmosphere, pressure p_0 . Assume the density constant. Find the expressions for the velocity of efflux and for the force exerted on the gas container due to the efflux. If the gas is oxygen at a pressure of 4 atmospheres in the tank, calculate the efflux velocity (1) with the density constant, and (2) taking into account the variation of density with pressure, assuming an adiabatic expansion.

8. With the help of Gauss's theorem prove the theorem of the last chapter that the stress tensor is symmetric.

9. Calculate the rate of discharge of a cylindrical pipe standing vertically, the liquid flowing in laminar flow under the action of gravity only.

10. A perfect gas at constant temperature is in equilibrium under the action of gravity. Find the relation between the pressure of the gas and the height above the surface of the earth.

11. Carry through the derivation of the laws of motion of viscous fluids using the modified form of Hooke's law and the general equations of motion of an elastic medium.

CHAPTER XVIII

HEAT FLOW

The problem of heat flow, although of quite different physical nature from elasticity and hydrodynamics, involves similar mathematics. Indeed, Fourier was concerned with problems of heat flow when he developed the series known by his name which we have used so much in our study of vibrations. First we set up the differential equation governing heat flow in a manner similar to the reasoning of the preceding chapters.

123. Differential Equation of Heat Flow.—The fundamental physical fact is that when there is a difference of temperature in a material body, heat will flow, and the rate of flow is proportional to the temperature gradient. Suppose we have a slab of thickness L , area a , with a difference of temperature $T_2 - T_1$ between the faces. Then the amount of heat flowing per second across the face is $\frac{-ka(T_2 - T_1)}{L}$, where k is the thermal conductivity,

the negative sign meaning that if $T_2 > T_1$, the flow will be backward toward low temperature. In the limit of an infinitely thin slab, this is simply $-ka\frac{\partial T}{\partial x}$, if x is the coordinate measured in the

direction of the heat flow. Next, there is the fact that if heat flows into a region, its temperature rises, the amount of rise being given by the relation that the amount of heat flowing in equals the change of temperature times the heat capacity, which in turn is the specific heat c times the mass. Putting these together, we obtain an equation which states the following: the rate of heat flow into a body is proportional to the time rate of change of its temperature; or, looking at it in another way, it is proportional to the temperature gradient around its boundaries. By eliminating the heat flow, we obtain a differential equation for the temperature.

Our first principle, which we have stated in the form that $-ka\frac{\partial T}{\partial x}$ measures the heat flow across the area a perpendicular to the x axis, is evidently a special case of the general law that

the flux density of heat flow is $f = -k \text{ grad } T$. This incidentally shows us at once that, if k is a constant, f is derivable from a potential, in this case kT , so that the curl of the flux is zero. The surfaces of constant temperature are called isothermals, and they serve as equipotentials, the lines of flow being at right angles to the isothermals. The equation of continuity now states that the time rate of increase of heat per unit volume equals the rate at which the heat flows in over the surface, plus the rate at which heat is produced inside. To raise the temperature of unit volume one degree requires an amount of heat equal to the heat capacity, or $c\rho$, if c is the specific heat, ρ the density of matter. Thus the time rate of increase of heat is $c\rho$ times the time rate of increase of temperature. We have then

$$\iiint c\rho \frac{\partial T}{\partial t} dv = - \iint f_n dS + \iiint P dv,$$

where P means the rate of production of heat per unit volume. By Gauss's theorem, the second term becomes $-\iiint \text{div } f dv$, so that for a small volume we have

$$c\rho \frac{\partial T}{\partial t} = -\text{div } f + P.$$

Substituting,

$$c\rho \frac{\partial T}{\partial t} = k \text{ div grad } T + P = k\nabla^2 T + P. \quad (1)$$

This is the equation of heat flow. At a point where heat is not being produced, it reduces to

$$\nabla^2 T = \frac{c\rho}{k} \frac{\partial T}{\partial t}, \quad (2)$$

an equation similar to the wave equation as far as the dependence on space is concerned. It contains, however, a first rather than a second time derivative, and this results in solutions which are exponentially damped, like a particle with resistance but no restoring force, rather than oscillating solutions. The particular case where the temperature is independent of the time, the steady state, leads simply to Laplace's equation, the term in time vanishing.

124. The Steady Flow of Heat.—The isothermals and lines of flow for the steady flow of heat are determined from Laplace's equation, and in some elementary cases we can find them with great ease. First let us consider a one-dimensional flow, which we

obtain with a slab of a substance, like a window pane, assuming that the temperature varies only with the coordinate x normal to the surface, being independent of y and z . Laplace's equation becomes $\partial^2 T / \partial x^2 = 0$, so that $T = a + bx$, with a constant temperature gradient. Thus if a face at $x = 0$ is kept at temperature T_0 , the other face at $x = L$ at T_1 , the temperature at intermediate points is given by $T = T_0 + (x/L)(T_1 - T_0)$. It is this simple case which furnishes the basis for the usual definition of thermal conductivity.

The cylinder forms a slightly more difficult problem in steady flow. For instance, let us ask for the steady state of temperature within a pipe formed of two concentric cylinders, whose inside and outside faces are kept at fixed temperatures. The temperature will depend only on r , and will be determined, on account of the divergenceless nature of the flow, by the condition that the same amount of heat flows across the surface of any cylinder with radius intermediate between r_0 and r_1 , the minimum and maximum radii of the pipe. This amount of heat is the product of the normal component of the flow, which is $f_r = -k(dT/dr)$, by the area of the cylinder, which for unit length along the pipe is $2\pi r$. In other words, $2\pi r f_r = -2\pi k r (dT/dr) = \text{constant}$, $dT/dr = a/r$, $T = a \ln r + b$. The two constants can be determined by fitting the temperatures at the two surfaces of the pipe. This example is interesting in showing that the temperature gradient is not always a constant in the steady state. The reason is very simple: the tubes of flow are not of constant cross-sectional area, and thus with a divergenceless flow the number of lines of flow per square centimeter, and consequently the magnitude of the temperature gradient and flux vector, must change from point to point. The same thing is evident in the flow of heat in a sphere, where the flow through concentric spheres must be the same. Hence, since the areas of these spheres increase proportionally to the squares of the radii, the temperature gradient must be inversely proportional to the square of the distance from the center, and the temperature inversely as the first power. These relations are just like those of the field and potential of a point charge in electrostatics, and as we shall later see, for just the same reason: both are solutions of Laplace's equation.

125. Flow Vectors in Generalized Coordinates.—Complicated problems in the steady flow of heat, as in hydrodynamics and electrostatics, are best approached by introducing curvilinear

coordinates, so that the boundaries of the bodies are expressed by coordinate surfaces, as with the cylinder and sphere. This suggests the formulation of the equation of steady flow, or Laplace's equation, in such general coordinates. Let the coordinates be q_1, q_2, q_3 and let them be orthogonal coordinates, so that the three sets of coordinate surfaces, $q_1 = \text{constant}$, $q_2 = \text{constant}$, $q_3 = \text{constant}$, intersect at right angles. Now let us move a distance ds_1 normal to a surface $q_1 = \text{constant}$. By doing so, q_2 and q_3 do not change, but we reach another surface on which q_1 has increased by dq_1 , which in general is different from ds_1 . Thus, with polar coordinates, if the displacement is along the radius, so that r is changing, $ds = dr$; but if it is along a tangent to a circle, so that θ is changing, $ds = r d\theta$. In general, we have

$$dq_1 = h_1 ds_1, dq_2 = h_2 ds_2, dq_3 = h_3 ds_3, \quad (3)$$

where in polar coordinates the h connected with r is unity, but that connected with θ is $1/r$. The first step in setting up vector operations in any set of coordinates is to derive these h 's, which can be done by elementary geometrical methods.

126. Gradient in Generalized Coordinates.—The component of the gradient of a scalar S in any direction is its directional derivative in that direction. Thus the component in the direction 1 (normal to the surface $q_1 =$

constant) is $\frac{dS}{ds_1} = h_1 \frac{\partial S}{\partial q_1}$. For instance, in polar coordinates, the r

component is $\frac{\partial S}{\partial r}$, and the θ com-

ponent $\frac{1}{r} \frac{\partial S}{\partial \theta}$.

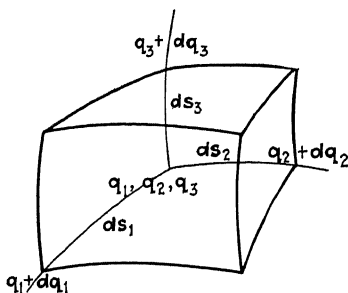


FIG. 32.—Element of volume for vector operations in curvilinear coordinates.

127. Divergence in Generalized Coordinates.—Let us apply Gauss's theorem to a small volume element $dV = ds_1 ds_2 ds_3$, bounded

by coordinate surfaces at $q_1, q_1 + dq_1$, etc. as in Fig. 32. If we have a vector A , of components A_1, A_2, A_3 along the three curvilinear axes, the flux into the volume over the face at q_1 , whose area is $ds_2 ds_3$, is $(A_1 ds_2 ds_3)_{q_1}$, and the corresponding flux out over the opposite face is $(A_1 ds_2 ds_3)_{(q_1 + dq_1)}$, where we note that the area $ds_2 ds_3$ changes with q_1 as well as the flux density A_1 .

Thus the flux out over these two faces is $\frac{\partial}{\partial q_1} (A_1 ds_2 ds_3) dq_1 =$

$\frac{\partial}{\partial q_1} \left(\frac{A_1}{h_2 h_3} \right) dq_1 dq_2 dq_3 = h_1 h_2 h_3 \frac{\partial}{\partial q_1} \left(\frac{A_1}{h_2 h_3} \right) dV$. Proceeding similarly with the other pairs of faces, and setting the whole outward flux equal to $\text{div } A \, dV$, we have

$$\text{div } A = h_1 h_2 h_3 \left[\frac{\partial}{\partial q_1} \left(\frac{A_1}{h_2 h_3} \right) + \frac{\partial}{\partial q_2} \left(\frac{A_2}{h_3 h_1} \right) + \frac{\partial}{\partial q_3} \left(\frac{A_3}{h_1 h_2} \right) \right]. \quad (4)$$

128. Laplacian.—Writing the Laplacian as $\text{div grad } \phi$, and placing $A_1 = \text{grad}_1 \phi$, etc., in the expression for $\text{div } A$, we have

$$\nabla^2 \phi = \text{div grad } \phi = h_1 h_2 h_3 \left[\frac{\partial}{\partial q_1} \left(\frac{h_1}{h_2 h_3} \frac{\partial \phi}{\partial q_1} \right) + \frac{\partial}{\partial q_2} \left(\frac{h_2}{h_3 h_1} \frac{\partial \phi}{\partial q_2} \right) + \frac{\partial}{\partial q_3} \left(\frac{h_3}{h_1 h_2} \frac{\partial \phi}{\partial q_3} \right) \right]. \quad (5)$$

It can easily be verified that this formula leads to the same values for the Laplacian in special cases which we have already obtained by direct differentiation in Chap. XV. But now we can understand the formula better, for we see that the terms like $h_1/h_2 h_3$ appearing inside the first differentiation arise from the fact that the flux through the opposite sides of a volume may differ not only on account of variation of the flux density, but also because the opposite sides can have different areas, as they do in the small volume element determined by coordinate surfaces with curvilinear coordinates.

129. Steady Flow of Heat in a Sphere.—Having obtained Laplace's equation in arbitrary coordinate systems, the problem of solving for the steady flow of heat becomes that of solving Laplace's equation in a suitable system, subject to certain boundary conditions. For instance, suppose we know that the surface of a sphere, radius r_0 , is kept at a temperature independent of time, though depending on the angles θ and ϕ . We then can set up the steady distribution of temperature within the sphere by solving Laplace's equation in spherical coordinates. The problem is mathematically like that of Problems 6, 7, and 8, Chap. XV, the vibration of a sphere, if we seek a solution independent of time. Just as in those problems, we separate variables in Laplace's equation, obtaining solutions of the form $\sin m\phi P_l^m(\cos \theta)R$, where the P 's are called associated Legendre polynomials, and where R satisfies the equation

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) - \frac{l(l+1)}{r^2} R = 0,$$

which can be immediately solved by setting $R = r^n$, where n is an integer to be determined. Substituting, this leads at once to the equation $n(n+1) = l(l+1)$, which has two solutions, $n = l$ or $n = -(l+1)$. In the present case, where the function must stay finite within the sphere, at $r = 0$, we cannot have inverse powers, so that the only allowable functions are r^l . Other problems solved by the same method, however, as for instance those of the electrostatic fields of distributions of charges, often involve functions which may become infinite at $r = 0$ but remain finite at large r 's, and they must be expanded in the series of inverse powers. We now have for a general solution

$$\sum_l \sum_m (A_{ml} \sin m\phi + B_{ml} \cos m\phi) P_l^m(\cos \theta) r^l.$$

To get the coefficients of the various terms in the sum, we set $r = r_0$, and determine the coefficients so that the resulting function of θ and ϕ is the assumed temperature distribution. This amounts to an expansion of the assumed function in series in the orthogonal functions $(\sin m\phi \text{ or } \cos m\phi) P_l^m(\cos \theta)$, and can be done by the usual methods for such expansions.

130. Spherical Harmonics.—To understand the physical meaning of the various terms of the expansion, we should consider the spherical harmonics, or functions of angles. Solving for these as in the problems quoted above, we find for the first few functions the following values:

$l = 0, m = 0$: constant

$l = 1, m = \pm 1$: $(\sin \phi \text{ or } \cos \phi) \sin \theta$

$m = 0$: $\cos \theta$

$l = 2, m = \pm 2$: $(\sin 2\phi \text{ or } \cos 2\phi) \sin^2 \theta$

$m = \pm 1$: $(\sin \phi \text{ or } \cos \phi) \sin \theta \cos \theta$

$m = 0$: $3 \cos^2 \theta - 1$.

These functions are shown graphically in Fig. 33, where the intersections of the nodal planes or cones with unit sphere are drawn. Thus the functions with $l = 1$ have one nodal plane, which may be perpendicular to any one of the three coordinate axes. This is seen most easily by remembering that $x = r \sin \theta \cos \phi$, $y = r \sin \theta \sin \phi$, $z = r \cos \theta$, so that the three solutions of the problem corresponding to $l = 1$ (r times the functions of angle) are simply x , y , z . These are obviously solutions of Laplace's equation, and have the nodal planes

$x = 0$, $y = 0$, $z = 0$, respectively. Similarly by making linear combinations of these three functions, we obtain solutions having any desired nodal plane. This is analogous to the degeneracy in the circular membrane, discussed in Sec. 103. With $l = 2$, there are two nodal surfaces, and so on. For discussing the vibrations of a sphere, of course these nodes would represent the regions of no displacement, the material on one side being displaced one way, the material on the other side in the opposite

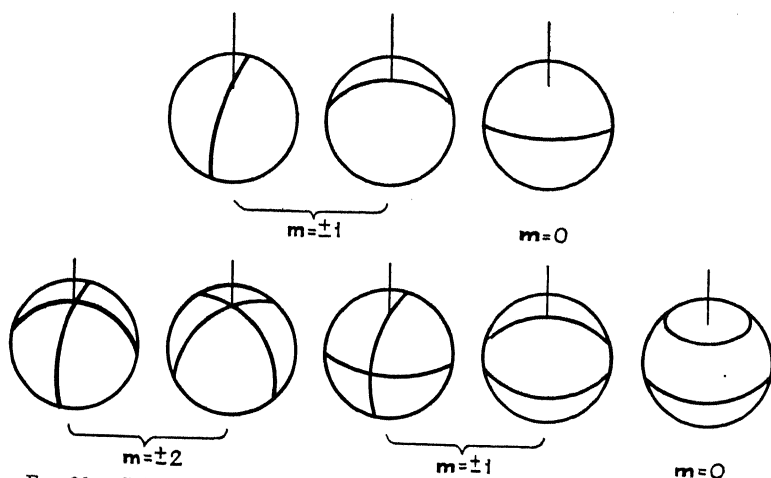


FIG. 33.—Spherical harmonics. Figures represent nodal lines on the surface of a sphere, for the functions $\sin m\phi P_l^m(\cos \theta)$ and $\cos m\phi P_l^m(\cos \theta)$. Upper line, $l = 1$; lower line, $l = 2$.

direction. With heat flow, the separate terms represent simple types of steady temperature distribution. For instance, the terms with $l = 1$ represent spheres in which the surface temperature varies as the cosine of the colatitude angle, or as the distance in a direction along the axis, and our solution tells us that in this case the temperature within the body varies linearly with distance, as in a flat slab. Higher terms represent more complicated solutions, and by superposing them any desired steady heat flow can be built up.

131. Fourier's Method for the Transient Flow of Heat.—The simplest type of problem in the transient flow of heat is the following: At $t = 0$, a body has a temperature which is an arbitrary function of position. At that instant, it is plunged into a cooling bath of some sort, which instantly cools its surfaces to a fixed distribution of surface temperature which is

maintained after that. The problem is to find the temperature throughout the body as a function of time as it cools from its initial to its final steady state. This can be easily reduced to a simpler case. We write the temperature at any time as the sum of two terms, the transient solution, and the steady-state solution. The latter is the temperature distribution set up by the cooling baths around the surface, and is discussed as in the last few sections in which steady flow of heat has been considered. The transient solution starts off with a temperature distribution which, added to the steady-state solution, gives the assumed initial temperature distribution of the body, and then gradually damps down to zero, finally leaving the steady-state solution only. Since at any instant after $t = 0$ the steady-state solution by itself gives the correct boundary temperature about the surface of the body, we see that the transient must give zero temperature at all points of the surface, independent of time. Thus the transient by itself is the solution of the problem in which a body is heated to an arbitrary temperature distribution at $t = 0$, after that is plunged into a cooling bath maintaining its whole surface at temperature zero, and gradually cools down to this temperature. We investigate this transient problem.

First we take the one-dimensional case, again of a slab, in which the initial temperature is an arbitrary function of x , but at all times after $t = 0$ the two faces, at $x = 0$ and $x = L$, are maintained at $T = 0$. The heat-flow equation becomes

$$\frac{\partial^2 T}{\partial x^2} = \frac{c\rho}{k} \frac{\partial T}{\partial t} = A \frac{\partial T}{\partial t}, \text{ if } A = \frac{c\rho}{k}.$$

We solve this equation by separation of variables. If $T = X(x)\Theta(t)$, and if we substitute in the equation and divide by T , we have

$$\frac{1}{X} \frac{d^2 X}{dx^2} = \frac{A}{\Theta} \frac{d\Theta}{dt} = -C^2.$$

Then separating we have

$$\frac{d\Theta}{dt} + \frac{C^2\Theta}{A} = 0, \quad \frac{d^2 X}{dx^2} + C^2 X = 0.$$

The solutions are

$$\Theta = e^{\frac{-C^2 t}{A}}, \quad X = \sin Cx \text{ or } \cos Cx.$$

We see that the temperature decreases exponentially with the time, approaching a constant value, a very reasonable behavior.

The boundary condition is now $T = 0$ when $x = 0$, $x = L$, and we satisfy this as we would with the vibrating string: we take only sines, and only those which reduce to zero at $x = L$; that is, we take $\sin (n\pi x/L)$, where n is an integer. In other words, $C = n\pi/L$, so that the function is constant $\times e^{-(n^2\pi^2/AL^2)t}$ $\sin (n\pi x/L)$, and the whole solution, writing in the value of A , is

$$\sum_n K_n e^{-\frac{n^2\pi^2 k}{c\rho L^2}t} \sin \frac{n\pi x}{L}. \quad (6)$$

Let us assume that the temperature distribution at $t = 0$ is $T = f(x)$. Then we wish to find the coefficients K_n , determining the temperature at later times. At $t = 0$ the exponentials

go to 1, so that we have $f(x) = \sum K_n \sin \frac{n\pi x}{L}$. We can then find

the coefficients K_n by Fourier's method, so that the problem is solved. The qualitative nature of the solution is easy to see. The original shape of the temperature curve will be distorted as time goes on, since the terms with high n damp down more rapidly than the others. After a certain lapse of time the whole slab will have become cooler, but also with a more simple temperature distribution, approximating the single term with $n = 1$. Thus, for instance, if it is originally all at a constant high temperature, and then is cooled, the original temperature curve would rise discontinuously from 0 at the edge to a constant value T inside. But after a time the curve would be like a single loop of a sine curve, showing that the edges would cool more rapidly than the middle.

The transient flow of heat in bodies of other shape may be considered by extensions of the same method. Thus the transient flow in the cylinder or sphere can be handled by introducing cylindrical or spherical polar coordinates, and separating variables just as for the vibration problems. The solutions, as far as the coordinates are concerned, come out as with vibrations, leading, for example, to sines and cosines of the angle, and Bessel's functions of r , in the case of two-dimensional flow in a circle or cylinder, but the time enters as a real exponential damping down to zero, rather than a complex exponential or sinusoidal function. Special cases are discussed in the problems.

132. Integral Method for Heat Flow.—There is another, different, method of great use in discussing the transient flow of heat.

This method is based on an important particular solution of the heat-flow equation. If we consider again the one-dimensional flow, and let $\alpha^2 = k/c\rho$, we can easily show that the function

$$f(x - x', t) = \frac{1}{2\alpha\sqrt{\pi t}} e^{-\frac{(x-x')^2}{4\alpha^2 t}} \quad (7)$$

is a solution of the equation, where x' is an arbitrary constant. To prove this, it is only necessary to substitute in the differential

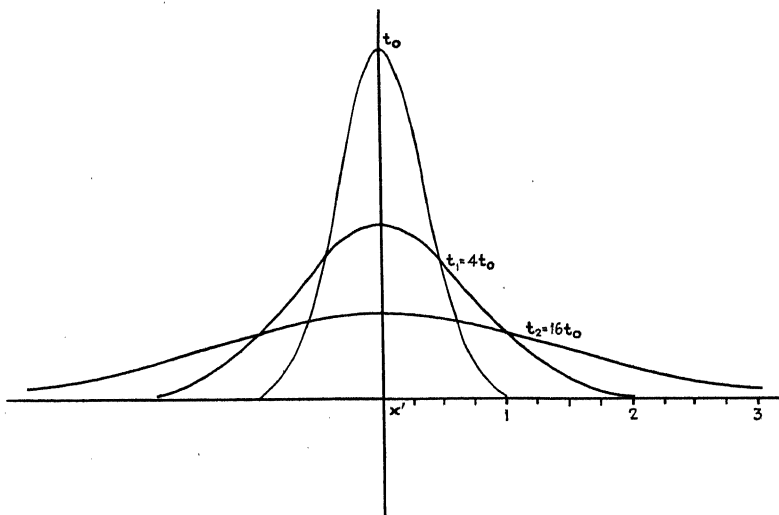


FIG. 34.—Function $f(x - x', t)$ of Eq. (7), as function of x , for different t 's. The function represents temperature distribution at different times resulting from initial conditions where the temperature is infinite at x' , zero elsewhere.

equation. The graph of the function f , plotted against x for different values of t , as in Fig. 34, has a sharp maximum at $x = x'$, looking like the familiar Gauss curve for probability distributions. At $t = 0$ the curve is coincident with the x axis everywhere except at $x = x'$, where it forms an infinitely high and narrow mountain, so that the area under the curve is finite. As time goes on, this mountain becomes flatter and broader, until finally the function is zero everywhere.

The function f can be used to discuss the following problem: At $t = 0$ the temperature throughout an infinite body is given by a function $T_0(x)$, and we are interested in the way in which this temperature distribution changes with time. We can break up the problem into a sum of other simpler problems, by dividing up

the distance x into small intervals, by a succession of points x_1, x_2, \dots, x_n . We set up the following problems:

1. The initial temperature is $T_0(x_0)$ between x_0 and x_1 , but is zero elsewhere;

2. The initial temperature is $T_0(x_1)$ between x_1 and x_2 , but is zero elsewhere;

...

n . The initial temperature is $T_0(x_{n-1})$ between x_{n-1} and x_n , but is zero elsewhere.

The initial temperature distribution connected with one of these problems would be similar to the curve of Fig. 34, for very small value of t , in that it would be large in a very small region, negligible or zero elsewhere. To make the maximum come at the right place, we must choose x' for the i th problem equal to x_i . As time goes on, the function f gives a good approximation to the way in which the temperature in this simple problem changes. Now if, at $t = 0$, we add together all the temperatures of Probs. 1 to n , we get the correct initial distribution of temperature. Therefore, if we add all the solutions at a later time, we again get the solution for the whole problem. This, of course, actually becomes an integral, the element of the integrand connected with the interval dx_i , which equals $x_{i+1} - x_i$, being proportional to $T_0(x_i)f(x - x_i, t)dx_i$. As a matter of fact, the constant of proportionality in f is so chosen that this gives just the right answer:

$$T(x, t) = \int_{-\infty}^{\infty} T_0(x') f(x - x', t) dx'. \quad (8)$$

To prove this, we need to do two things: first, prove that it is a solution of the heat-flow equation; secondly, show that it approaches the correct value at $t = 0$. The first is obvious, for the integrand, regarded as a function of x and t , has already been shown to be a solution of the equation, and on account of the linear nature of the differential equation a sum of solutions is a solution. For the second, we note that at $t = 0$ the function $f(x - x', t)$ has appreciable values only at $x = x'$. The whole integral will then come from the immediate neighborhood of $x' = x$, so that we may insert this value in T_0 , and take it outside the integral sign, obtaining

$$T(x, 0) = T_0(x) \int_{-\infty}^{\infty} f(x - x', 0) dx'.$$

The integral is $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-u^2} du$, where $u = \frac{x' - x}{2a\sqrt{t}}$, and this equals unity. Hence we have shown that $T(x, 0) = T_0(x)$, so that we have verified our solution.

By a slight variation, it is possible to solve the problem in which the temperature of a semi-infinite slab bounded by $x = 0$ is initially any desired value, and in which the surface is kept at $T = 0$ at all subsequent times. Let the initial temperature be $T_0(x)$, where this function is defined only for positive x 's, inside the slab. We now define an odd function equal to $T_0(x)$ for positive x 's, equal therefore to $-T_0(-x)$ for negative x 's. If we set up an infinite slab with this temperature distribution, then on account of symmetry the temperature at $x = 0$ will always be zero, and our boundary condition is satisfied, the part of the solution for positive x 's being the desired function.

Integral methods similar to that described can be used also to discuss the problem in which the surface of a semi-infinite slab is kept at a temperature which varies in an arbitrary way with time. Two- and three-dimensional problems can also be treated, though the principles are not essentially different from those already considered.

One interesting feature of heat flow is brought out by the integral solution which we have just used. That is its irreversible nature. Thermodynamically, heat conduction is a typical irreversible process, and this is shown in the fact that heat always flows from the warmer to the cooler body, never in the opposite direction. With reversible processes, as for instance vibration problems, one can change the sign of the time where it appears in the solution and still have a possible solution of the equation; a vibration running backward is not essentially different from one running forward. But that is not the case in the heat-flow equation, as we see easily from Eq. (7), where, if we attempt to give t a negative value, the solution becomes imaginary. The essential mathematical difference between the two cases is that in heat flow a first time derivative appears, while in vibration problems and wave equations there is a second time derivative. This second time derivative is unchanged when t is changed to $-t$, whereas the first time derivative in the heat-flow equation changes sign with t , so that, if a given function satisfies the equation, it will no longer satisfy it if time is reversed.

Problems

1. Derive the divergence, gradient, and Laplacian in spherical polar coordinates by the general method of this chapter.
2. Discuss the steady flow of heat in a spherical shell contained between two concentric spheres, the temperature being an arbitrary function of position over both surfaces.
3. Discuss the steady two-dimensional flow of heat in a semi-infinite rectangular bar bounded by $x = 0$, $x = L$, $y = 0$, extending to infinity along the y axis, subject to the boundary condition that the temperature is zero along the two infinite sides of the bar, but that it is an arbitrary function of x along the end from $x = 0$ to $x = L$. Build up the solution out of individual solutions varying sinusoidally with x , and exponentially with y , noting that they must decrease rather than increase exponentially as y increases.
4. Discuss the steady flow of heat in a semi-infinite cylindrical rod with a flat end, if the temperature is kept at zero along the cylindrical face, but is an arbitrary function of position on the end.
5. A slab is heated to a uniform temperature T_1 , then plunged in a bath which keeps its temperature at T_0 . Find the interior temperature as a function of the time, computing and drawing several graphs, so chosen as to show the progress of the cooling process.
6. For small times after the cooling process has commenced in Prob. 5, the interior temperature will not have changed appreciably, and the slab will act practically like a semi-infinite slab. Compare the solution of Prob. 5, using Fourier's method, with the corresponding solution by the integral method, computing both curves and comparing.
7. In an infinite body the temperature is initially unity between the planes $x = -1$ and $x = 1$, and is zero everywhere else. Plot the temperature as a function of x for several instants of time, and finally for $t = \infty$. (Use Peirce's tables for values of the Gauss error function $\int_u^\infty e^{-u^2} du$.)
8. Prove that the integral $\int_0^\infty e^{-u^2} du = \frac{\sqrt{\pi}}{2}$. (Suggestion: Multiply this integral by the equal integral $\int_0^\infty e^{-v^2} dv$, and consider u and v as Cartesian coordinates in a plane. Introduce polar coordinates in the plane, carrying out the integration in those coordinates.)
9. Show that a particular integral of the equation for heat flow in an infinite medium is $\frac{\text{constant}}{t^{\frac{3}{2}}} e^{-\frac{r^2}{4a^2t}}$, where r is the distance from the origin. Discuss the initial temperature distribution corresponding to this solution.
10. Show that the integral

$$T = \frac{1}{(2a\sqrt{\pi t})^3} \iiint e^{-\frac{r^2}{4a^2t}} T_0(x'y'z') dx' dy' dz'$$

is a general solution of the heat-flow equation in three dimensions corresponding to an initial temperature distribution of $T_0(x, y, z)$, where $r^2 = (x - x')^2 + (y - y')^2 + (z - z')^2$.

CHAPTER XIX

ELECTROSTATICS, GREEN'S THEOREM, AND POTENTIAL THEORY

The problems of electrostatics are practically identical mathematically with those of flow, which we have been considering in the last few chapters. The fundamental physical law is very simple. Electric charges exert forces on each other, given by Coulomb's law, which states that the force is directed along the line of centers, and equal to ee'/r^2 , where e and e' are the strengths of the charges, r the distance between. The force on a particular charge is then given as the sum of the individual attractions and repulsions exerted by all the other charges. The force per unit charge at any point is the intensity of the electric field, a vector function of position. The lines tangent to the force vector, similar to the lines of flow in the last two chapters, are called the lines of force.

133. The Divergence of the Field.—Consider the field of a point charge at the origin of coordinates. The field intensity E is a vector of magnitude e/r^2 , pointing out along the radius; its components are thus

$$\frac{ex}{r^3}, \frac{ey}{r^3}, \frac{ez}{r^3}.$$

We then have

$$\begin{aligned} \operatorname{div} E &= \frac{\partial}{\partial x} \left(\frac{ex}{r^3} \right) + \frac{\partial}{\partial y} \left(\frac{ey}{r^3} \right) + \frac{\partial}{\partial z} \left(\frac{ez}{r^3} \right) = \\ &= e \left[\frac{3}{r^3} - \frac{3(x^2 + y^2 + z^2)}{r^5} \right] = 0 \end{aligned}$$

We thus see that the field of a point charge is divergenceless. In other words, if we represent the field strength by the number of lines of force per square centimeter, these lines will never start or stop in empty space. They will, of course, start or stop on charges. We cannot see this directly, but we can prove it by using Gauss's theorem. Take a small sphere of radius R about the origin. Then we know that the volume integral of the divergence of E over the volume equals the surface integral

of the normal component of E . This component is e/R^2 , and the surface area is $4\pi R^2$, so that the surface integral in question is $4\pi e$. Thus the volume integral of the divergence over our small volume is $4\pi e$, which is different from zero. Since the number of lines emerging across an area equals the field strength, the total number of lines of force diverging from the charge e is also $4\pi e$.

Now consider the field of many point charges. The field of each charge separately has zero divergence. Therefore, since the divergence of the sum of several functions is the sum of the divergences, it is plain that the divergence of the whole field vanishes: $\text{div } E = 0$ in general. The only exception is for those points where there is charge, for there we have seen that the divergence does not vanish. Let us see what does happen there. In the first place we introduce ρ , the volume density of charge. Now take a small volume dv , containing a charge ρdv . Surely if dv is small enough this field will be just as if the same charges were concentrated at a point. Thus $4\pi\rho dv$ lines will diverge from the charge, or $\iint E_n dS = \text{div } E dv = 4\pi\rho dv$. Dividing by dv , we have

$$\text{div } E = 4\pi\rho. \quad (1)$$

This is the general equation for the divergence of the field, and we see that it reduces to $\text{div } E = 0$ at points where the charge density vanishes. This equation, $\text{div } E = 4\pi\rho$, is mathematically equivalent to the continuity equation

$$\frac{\partial \rho}{\partial t} = -\text{div } f + P,$$

if we set the time derivative equal to zero, and consider $4\pi\rho$ as the quantity analogous to the rate of production of material. Here, of course, there is no actual idea of flow, the analogy being merely mathematical.

134. The Potential.—We can immediately show that the curl of the field of a point charge vanishes. And unlike the divergence equation, this is true everywhere, even right at the charge. Then, if we superpose many charges, the curl still is zero, so that we have the general equation $\text{curl } E = 0$. This holds in all static cases (we shall later have a term to add to the equation, containing a time derivative). Thus we can always set up an electrostatic potential ϕ , such that $E = -\text{grad } \phi$. Taking the divergence, we find the equation which the potential satisfies: it is

$$-\operatorname{div} \operatorname{grad} \phi = -\nabla^2 \phi = 4\pi\rho, \quad (2)$$

which is called Poisson's equation. Laplace's equation $\nabla^2 \phi = 0$ is the special case which holds in those regions of space that contain no charge.

If we form the line integral of the electric field intensity along a given curve between two points of the field, A and B , then $\int E \cdot ds$ along this curve is called the electromotive force along the path. It is obviously the work per unit charge done by the field when a charge is moved along the given path from A to B . In the electrostatic case, since E can be obtained from a potential, $E = -\operatorname{grad} \phi$ and

$$\begin{aligned} \text{E.m.f.} &= \int_A^B E \cdot ds = - \int_A^B \operatorname{grad} \phi \cdot ds = \\ &= - \int_A^B \left(\frac{\partial \phi}{\partial x} dx + \frac{\partial \phi}{\partial y} dy + \frac{\partial \phi}{\partial z} dz \right) \\ &= - \int_A^B d\phi = \phi_A - \phi_B \end{aligned}$$

so that in this case the e.m.f. is equal to the potential difference between the points A and B . The distinction between e.m.f. and potential difference is of importance in cases where $\operatorname{curl} E \neq 0$ and hence there is no potential. Even in this case we may still use the idea of e.m.f.

135. Electrostatic Problems without Conductors.—There are two principal sorts of electrostatic problems. The first is that in which we know the distribution of charge, and wish to compute the field. We could always do this by direct summation of the fields due to the individual charges, but often that is very difficult, and we can simplify greatly by using the potential and Laplace's equation. Thus suppose we have charge uniformly distributed over an infinite plane, the amount per unit area being σ , and suppose we wish the field at a distance R from that plane. We may get this by a direct calculation. Thus we take a set of polar coordinates in the plane, which center at the point directly beneath the place where we wish the potential, as in Fig. 35. Between the circles of radius r and $r + dr$, and between θ and $\theta + d\theta$, will be an amount of charge $\sigma r d\theta dr$. This will be at a distance $\sqrt{R^2 + r^2}$ from the point we are interested in, so that its field will have the magnitude $\frac{\sigma r d\theta dr}{R^2 + r^2}$. The component normal

to the plane, which is all that we need, is this times $\frac{R}{\sqrt{R^2 + r^2}} =$

$\frac{\sigma R r d\theta dr}{(R^2 + r^2)^{3/2}}$. The total field is then

$$\int_0^{2\pi} d\theta \int_0^\infty \frac{\sigma R r dr}{(R^2 + r^2)^{3/2}} = 2\pi\sigma \int_0^\infty \frac{x dx}{(1 + x^2)^{3/2}},$$

where $x = \frac{r}{R}$.

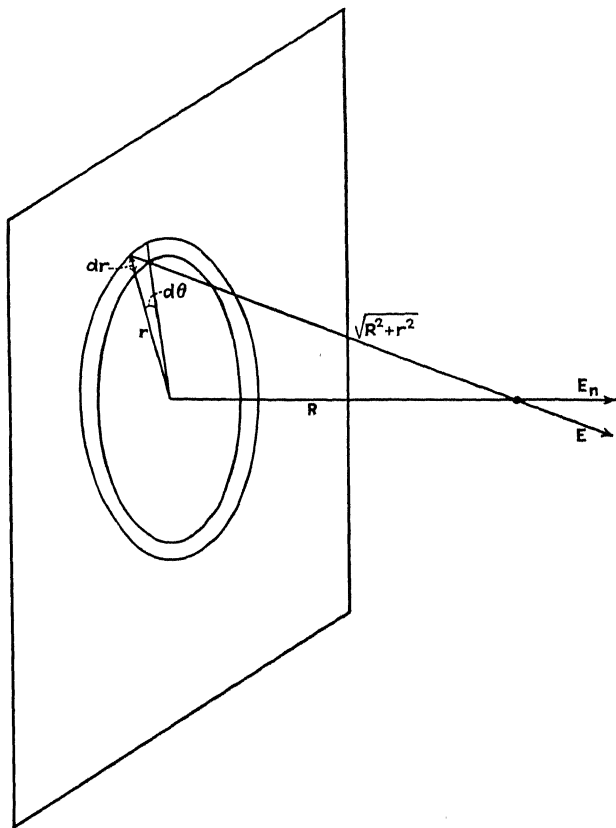


FIG. 35.—Field of a charged plane. From charge between r and $r + dr$, θ and $\theta + d\theta$:

$$E = \frac{\sigma r d\theta dr}{R^2 + r^2}, \quad E_n = \frac{\sigma R r d\theta dr}{(R^2 + r^2)^{3/2}}.$$

Letting $1 + x^2 = y$, so that $x dx = dy/2$, this is

$$\frac{2\pi\sigma}{2} \int_1^\infty \frac{dy}{y^{3/2}} = \frac{2\pi\sigma}{2} (-2y^{-1/2}) \Big|_1^\infty = 2\pi\sigma.$$

Thus the field is a constant, independent of position. Similarly on the other side of the plane it is $-2\pi\sigma$, so that there is a discontinuity in E of $4\pi\sigma$ in crossing the surface.

We have seen that it is possible in such a simple case to compute the field directly. But it is done far more easily by using our general principles. Thus the potential can depend only on the coordinate normal to the plane, which we denote by x . Its differential equation, outside the charged sheet, is then

$$\frac{d^2\phi}{dx^2} = 0,$$

$$\phi = ax + b,$$

showing that the field is constant everywhere, and in the x direction. To investigate conditions on the surface, we set up a thin

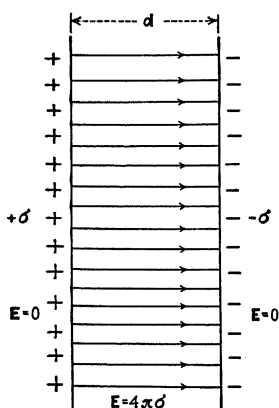


FIG. 36.—Field in parallel plate condenser.

flat volume, with its broad sides parallel to the charged plane, and enclosing just 1 sq. cm. of this plane. It will then hold charge σ , so that $4\pi\sigma$ lines will diverge from it. By symmetry, these will leave it at right angles, and an equal number over each face. Hence $2\pi\sigma$ will leave over each face, or the field strength is $2\pi\sigma$ on the one side, $-2\pi\sigma$ on the other. We have the same result as before, with very much simpler calculation.

Similar problems are met in the theory of the condenser. Take, for example, the parallel plate condenser, as in Fig. 36, two charged plates of area A , so large in proportion to their separation d that they can be almost treated as infinite. Let the charge per square centimeter be σ on one plate, $-\sigma$ on the other. Then we must find the potential difference between the plates, for by definition the capacity $C = \frac{\sigma A}{PD}$. But now, just as in the last case, the field must be constant and perpendicular to the plates. It can have different values in the three regions to the left of the plates, between, and to the right. And it has a discontinuity of $4\pi\sigma$ in passing through a plate of surface density σ . These conditions are all satisfied by having no field outside the condenser, and by having a field $4\pi\sigma$ within, pointing from the positive plate to the negative. Thus

the potential difference, being the field times the distance, is $4\pi\sigma d$, so that

$$C = \frac{\sigma A}{4\pi\sigma d} = \frac{A}{4\pi d}, \quad (3)$$

the familiar formula for a parallel plate condenser. It should be noticed that capacitance has the dimensions of a length in the electrostatic system of units.

136. Electrostatic Problems with Conductors.—The second sort of electrostatic problem is more difficult. It is that in which there are conductors as well as charges. Now in the presence of a charge, induced charges are set up on conductors, and it is usually a difficult problem to find how they are distributed, and hence to find their field. In this case it is practically indispensable to make use of the methods of potential theory. To see how to proceed, let us imagine the train of events which would occur when a charge was brought near a conductor. The charge would carry with it a field, which in general would be such that different parts of the conductor were at different potentials. Now a conductor has the peculiarity that if there is a field in it, a current flows, and continues to flow as long as the field remains. Thus charge will start to flow through the conductor, being attracted or repelled by the external charge. This will continue until just such a charge distribution has been set up in the conductor that the field resulting from it plus the external charges reduces to zero within the conductor, or the potential throughout the conductor is constant, for this is the condition for no current flow. In other words, the whole of a conductor, surface and inside, is part of a single equipotential. We then solve such a problem in the following way: we look for a solution of Poisson's equation, holding in the region outside the conductors, and reducing to constants on the boundaries. This solution thus gives the potential of the problem, and its gradient gives the field.

We can illustrate better by a problem. Consider an infinite conducting plane, uncharged as a whole, with a charge e in front of it at a distance d . Now we wish a solution of Poisson's equation, reducing to a constant over the face of the plane. We set this up by a device, called the method of images. We imagine the plate removed, its face replaced by an imaginary plane, and at a distance d behind the plane we put a charge $-e$, as if it were e 's image in a mirror, as shown in Fig. 37. Then these two

charges together would keep the whole plane just at potential zero. For any point of the plane is equidistant from both charges, one has the potential e/r , and the other $-(e/r)$, and they just cancel. The potential at any point of space can be easily found, now, in the field of these charges. It is simply

$$e\left(\frac{1}{r_1} - \frac{1}{r_2}\right),$$

if r_1 is the distance from the charge e , r_2 the distance from its mirror image. The lines of force and equipotentials look like

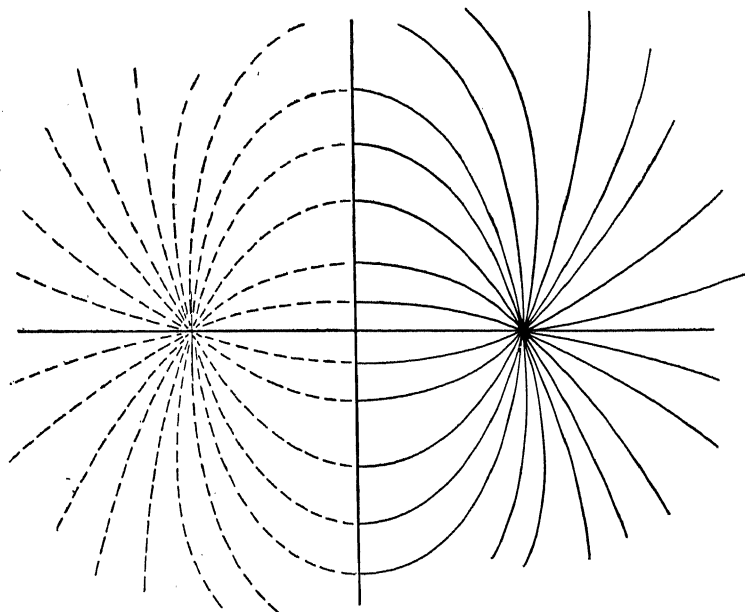


FIG. 37.—Lines of force for charge e in front of conducting plane, by method of images.

those of a bar magnet, and it is perfectly true that the plane bisecting the magnet is an equipotential. In our actual problem, now, the potential in the empty space is just that given by our field of two charges; in the metal the potential is zero.

We might naturally inquire what induced distribution of charge would be set up in the conducting plane, to produce this final field. In the first place, in a steady state, the charge within a conductor is always zero. For the field is zero within it, therefore its divergence is zero. Thus all charge is concentrated on the surface. Next, as we showed before, the normal component

of the electric field has a discontinuity of $4\pi\sigma$ at a surface carrying a surface charge σ . Thus if we can compute the discontinuity, we can in turn get the surface density of charge. In our case the field is normal to the plate, by symmetry, so that the discontinuity of E_n in crossing the surface is just equal to the total E outside. This may be found at once from our known potential function, so that we could get the necessary surface charge.

137. Green's Theorem.—The fundamental theorem of potential theory is a mathematical relation called Green's theorem. It is a result of Gauss's theorem, and is easily proved. Gauss's theorem states that $\iiint \operatorname{div} E \, dv = \iint E_n \, dS$ for any vector E . Now let $E = \phi \operatorname{grad} \psi$, where ϕ and ψ are two scalar functions, then $\operatorname{div} E = \operatorname{div} (\phi \operatorname{grad} \psi) = \phi \nabla^2 \psi + \operatorname{grad} \phi \cdot \operatorname{grad} \psi$, as we can easily prove. Also $E_n = \phi \frac{\partial \psi}{\partial n}$, where $\frac{\partial \psi}{\partial n}$ is the normal derivative, the component of the gradient along n . Hence we have

$$\iiint (\phi \nabla^2 \psi + \operatorname{grad} \phi \cdot \operatorname{grad} \psi) \, dv = \iint \phi \frac{\partial \psi}{\partial n} \, dS. \quad (4)$$

This is one form of Green's theorem. To get the more familiar form, we next write just the same expression with ϕ and ψ interchanged:

$$\iiint (\psi \nabla^2 \phi + \operatorname{grad} \phi \cdot \operatorname{grad} \psi) \, dv = \iint \psi \frac{\partial \phi}{\partial n} \, dS.$$

Now we subtract, obtaining

$$\iiint (\phi \nabla^2 \psi - \psi \nabla^2 \phi) \, dv = \iint \left(\phi \frac{\partial \psi}{\partial n} - \psi \frac{\partial \phi}{\partial n} \right) \, dS. \quad (5)$$

This is the common form of Green's theorem. We shall now consider a number of applications of this mathematical theorem. These applications come mostly in the discussion of methods of solving Poisson's and Laplace's equations. Of course, these can be solved by the method of separation of variables, and development in series of orthogonal functions. But the present method, called Green's method, is quite different, and almost more useful in a general discussion, though perhaps not in particular problems.

138. Proof of Solution of Poisson's Equation.—We can easily see how to solve Poisson's equation, $\nabla^2 \phi = -4\pi\rho$. For this gives the potential ϕ due to a charge distribution. Now if we

divide space into small elements of volume dv , the charge ρdv will exert a potential $\rho dv/r$, if r is the distance from the point where we wish the potential due to dv . Thus the whole potential is $\iiint \frac{\rho dv}{r}$. But $\rho = -\frac{1}{4\pi} \nabla^2 \phi$, so that we have

$$\phi = -\frac{1}{4\pi} \iiint \frac{\nabla^2 \phi}{r} dv, \quad (6)$$

giving the solution of Poisson's equation. In this integral, we must integrate over all space, so as to include all charges. We have derived our solution rather intuitively from the known solution for a point charge. But we can derive it rigorously from Green's theorem.

In the last form of Green's theorem, let $\psi = 1/r$, where r is the distance from a point P , and let ϕ be the potential ϕ . Thus we have

$$\iiint \left[\phi \nabla^2 \left(\frac{1}{r} \right) - \frac{1}{r} \nabla^2 \phi \right] dv = \iint \left[\phi \frac{\partial(1/r)}{\partial n} - \frac{1}{r} \frac{\partial \phi}{\partial n} \right] dS.$$

This is true no matter what volume we use. Let us choose as our volume the whole of space, except for a tiny sphere of radius R surrounding the point P where we wish to compute the potential. Now $\nabla^2(1/r) = 0$, except where $r = 0$, so that it is zero throughout the whole of our volume, and the left side becomes

$$-\iiint \frac{\nabla^2 \phi}{r} dv.$$

Let us compute the right side. The integral is to be taken over the surface of our volume, which consists of our tiny sphere, and a surface at infinity, which for the present we neglect. Over the surface of the tiny sphere, the direction n is simply the radial direction, pointing in toward P (because it is directed out of the volume). We have

$$\frac{\partial(1/r)}{\partial n} = -\frac{d(1/r)}{dr} = \frac{1}{r^2}, \quad \frac{\partial \phi}{\partial n} = -\frac{\partial \phi}{\partial r}.$$

Then the right side is

$$\iint \frac{\phi}{r^2} dS + \iint \frac{1}{r} \frac{\partial \phi}{\partial r} dS.$$

But on the surface of the sphere, $r = R$, so that this is

$$\frac{1}{R^2} \iint \phi dS + \frac{1}{R} \iint \frac{\partial \phi}{\partial r} dS.$$

Now $\frac{\iint \phi dS}{\iint dS}$ is just the mean value $\bar{\phi}$ of ϕ over the surface, and

$$\frac{\iint \frac{\partial \phi}{\partial r} dS}{\iint dS}$$

is the mean value of $\frac{\partial \phi}{\partial r}$. But the $\iint dS$ is the area of the sphere $= 4\pi R^2$, so that our integral is $4\pi \bar{\phi} + 4\pi R \frac{\partial \phi}{\partial r}$, and the whole relation is, changing sign,

$$\iiint \frac{\nabla^2 \phi}{r} dv = -4\pi \bar{\phi} - 4\pi R \frac{\partial \phi}{\partial r}.$$

If now R approaches zero, the last term vanishes, and $\bar{\phi}$ approaches ϕ , the value at the point P . Hence we have

$$\phi = -\frac{1}{4\pi} \iiint \frac{\nabla^2 \phi}{r} dv,$$

the solution of Poisson's equation which we wished to prove.

There are several points to be mentioned in connection with this proof. In the first place, the volume integral is taken over all space, except an infinitely small sphere surrounding P : a point charge exerts an effect on all other charges, but not on itself. Secondly, we neglected entirely the fact that our volume has a surface at infinity, which we should take into account in calculating our surface integrals. Suppose that the volume were not really infinite, but merely very large, being bounded, say, by a second large sphere of radius R' . Then the surface integral over the large sphere is similar to that over the small one, but with opposite sign: it is $4\pi \bar{\phi}' + 4\pi R' \frac{\partial \phi'}{\partial r'}$, where now $\bar{\phi}'$ is the mean over the large sphere, etc. To neglect these terms, as we have done, their limits must be zero as R' becomes infinite.

That is, $\bar{\phi}'$ must go to zero at infinite distance, and $R' \frac{\partial \phi'}{\partial r'}$ must also go to zero. These are both satisfied if ϕ is the potential of a set of charges at finite points, for then ϕ will go as $1/r$, $\partial \phi / \partial r$ will go as $1/r^2$, and $r \partial \phi / \partial r$ will fall off as $1/r$, becoming zero as r becomes infinite.

139. Solution of Poisson's Equation in a Finite Region.—

Suppose now that instead of extending our integral over all space, we integrate only over a finite volume V , with surface S , excluding in each case our infinitesimal sphere of radius R . Then plainly we have

$$-\iint\int\frac{\nabla^2\phi}{r}dv=4\pi\phi+\iint\left[\phi\frac{\partial(1/r)}{\partial n}-\frac{1}{r}\frac{\partial\phi}{\partial n}\right]dS,$$

or

$$\phi=-\frac{1}{4\pi}\iint\int\frac{\nabla^2\phi}{r}dv-\frac{1}{4\pi}\iint\left[\phi\frac{\partial(1/r)}{\partial n}-\frac{1}{r}\frac{\partial\phi}{\partial n}\right]dS, \quad (7)$$

where the volume integral is taken over the whole volume V , excluding the infinitesimal sphere, and the surface integral is taken over S .

We can explain this important formula in words much better than by mathematics. The potential at a given point can be written as the sum of two parts: the potential of all the charges within a certain finite volume surrounding the point, and another part, which, of course, must represent the potential of the other charges outside our volume. But the second term appears as a surface integral, not a volume integral. This is an example of the usual sort of application of Green's theorem: the replacement of a volume integral by a surface integral.

There is one interesting way of regarding the solution. Suppose first that ρ were zero all through our volume, though not outside. Then $\nabla^2\phi$ will be zero inside, and the volume integral will vanish. Further, ϕ will satisfy Laplace's equation within the region. The surface integral, in other words, represents a solution of Laplace's equation within our region, in terms of an integral over the boundary of the region. As a matter of fact, any solution of Laplace's equation in this region can be written in this way, by using the proper boundary values of ϕ and $\partial\phi/\partial n$ at the surface. The last two terms in our solution, in other words, represent a general solution of the homogeneous equation $\nabla^2\phi=0$, the arbitrary functions (which with partial differential equations replace the arbitrary constants) being the boundary values of ϕ and $\partial\phi/\partial n$. The volume integral, on the other hand, represents a particular solution of the inhomogeneous equation $\nabla^2\phi=-4\pi\rho$, satisfying the equation but not its boundary values. Thus we have the familiar case in which the solution of an inhomogeneous equation is the sum of a particular solution,

and the general solution of the related homogeneous equation. And this general solution is to be so chosen that the sum of both terms satisfies the boundary values, on the surface of the volume.

140. Green's Distribution.—When we examine the surface integral of Eq. (7) more in detail, we can see what it represents.

The term $\frac{1}{4\pi} \int \frac{1}{r} \frac{\partial \phi}{\partial n} dS$ represents evidently the potential arising

from a certain surface charge, of surface density $\frac{1}{4\pi} \frac{\partial \phi}{\partial n}$. The

other term, $-\frac{1}{4\pi} \int \int \frac{\phi \partial(1/r)}{\partial n} dS$, is a little complicated. The

term $\frac{\partial(1/r)}{\partial n}$ is the difference between the potential of two unit

charges, spaced at a distance dn along the normal, divided by dn ; that is, it is the potential of two charges, one of strength $1/dn$, the other $-1/dn$, at distance dn ,

as in Fig. 38. Such a combination of an equal and opposite positive and negative charge very close together is called a dipole. The strength of a dipole, or the dipole moment, is the strength of one of the charges times the distance of separation. Thus in our case the strength is $(1/dn)dn$, so that we have the potential of a unit dipole. Then the integral is the potential of a dipole distribution of moment $\phi/4\pi$ per unit area. Such a distribution is called a

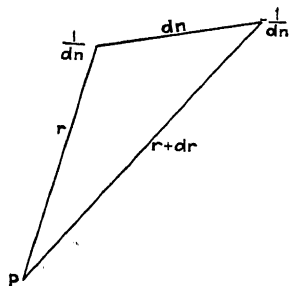


FIG. 38.—Potential of unit dipole, consisting of charges $\pm \frac{1}{dn}$ at distance dn apart.

double layer, since it consists of layers of positive and negative charges close together. We then see that by spreading on the surface of our region a suitable layer of surface charge, and a double layer of dipoles, we produce just the same field inside that the external charges would give. This distribution of charge and double layer is called Green's distribution.

Suppose that we know that a given function ϕ satisfies Laplace's equation within a given region. Suppose further that we know its boundary value ϕ , and its normal derivative $\partial \phi / \partial n$, at all points of the surface of the region. Then we can at once write the solution of Laplace's equation having these boundary values. It is

$$\phi = -\frac{1}{4\pi} \int \int \left(\phi \frac{\partial(1/r)}{\partial n} - \frac{1}{r} \frac{\partial \phi}{\partial n} \right) dS,$$

integrated over the boundary. This is obviously a very simple way of getting a solution of a differential equation satisfying given boundary values. In particular it is simpler than the methods we have used so far, in that we can apply it to any form of surface.

There is a simple interpretation of Green's distribution. Suppose that the field within our volume were just what it is, but that outside the volume the field and potential were everywhere zero. Then at the boundary there would be a discontinuity of potential and field. Now we have already seen that at a surface charge σ there is a discontinuity of field, $4\pi\sigma$, so that at a discontinuity of the field there is a surface charge equal to $1/4\pi$ times the discontinuity of the normal component of the field. Thus if the normal component of the field is zero outside, $\partial\phi/\partial n$ inside, the surface charge is $1/4\pi \partial\phi/\partial n$. This is just the surface charge concerned in Green's distribution. Similarly, at a boundary where there is a discontinuity of potential, there must be a double layer, of moment per unit area equal to $1/4\pi$ times the discontinuity of the potential, as we see from a condenser of charge σ , dipole moment σd per unit area, potential difference $4\pi\sigma d$. This gives the double layer of Green's distribution. In other words, these surface charges and layers, plus the charges within the region, are just those necessary to give the potential its actual values within the volume, and to reduce it to zero outside.

141. Green's Method of Solving Differential Equations.—

We have seen in the present chapter a method, called Green's method, for solving differential equations, quite different from any we have met before, except the integral method of treating heat flow, which is very similar. The most characteristic part of the method is in the solution of Poisson's equation, as an integral of ρ/r over all space. Here we had an inhomogeneous equation, $\nabla^2\phi = -4\pi\rho$. Suppose we let $\rho = \rho_1 + \rho_2 + \rho_3 \dots$, where ρ_i is equal to ρ in the i th volume element dv_i , but is zero elsewhere. Then we can write the equations $\nabla^2\phi_1 = -4\pi\rho_1$, $\nabla^2\phi_2 = -4\pi\rho_2$, \dots , for each of these, where ρ_i is different from zero only in a very small region, so that the problem is practically that of a point charge, which we can solve. We add these func-

tions to get the whole solution, according to Sec. 26, Chap. IV. This is the essence of Green's method, the separation of the inhomogeneous part of the equation into simple parts, each of which we can solve. The function $1/r$, which is the solution for one of these problems, is called the Green's function. As a matter of fact, a general method of solving differential equations by means of Green's functions has been worked out, and it lies at the basis of much of the more advanced work on the theory of differential equations, particularly of the second order.

Problems

1. Given a spherical distribution of charge, in which the density is a function of r . Prove that the field at any point is what would be obtained by imagining a sphere drawn through the point, with its center at the origin, all the charge within the sphere concentrated at the center, and all the charge outside removed. Apply to gravitation, showing that the earth acts on bodies at its surface as if its mass were concentrated at the center.

2. Given a sphere filled with charge of constant density. Prove that at points within the sphere, the field is directly proportional to the distance from the center.

3. A condenser consists of two concentric spheres, holding equal and opposite charges. Find its capacity. Similarly find the capacity of a condenser consisting of two long concentric circular cylinders.

4. Compute the surface density induced by a charge on a plane conductor.

5. In a certain spherical distribution of charge, the potential is given by $\frac{-e^{-ar}}{r}$. Find the charge density as a function of r . Also find the charge contained between r and $r + dr$. This represents roughly the charge distribution within an atom.

6. Prove $\text{div} (\phi \text{ grad } \psi) = \phi \nabla^2 \psi + \text{grad } \phi \cdot \text{grad } \psi$.

7. There are certain charges and conductors in an electrostatic field, whose potential is ϕ . Show that the surface density of charge on the surface of a conductor is $\frac{-1}{4\pi} \frac{\partial \phi}{\partial n}$, where n is the normal pointing out of the conductor.

Show that the electric field is normal to the surface of a conductor.

8. It requires several volts energy to remove an electron from the interior of a metal to the region outside. Find how many volts, if the double layer at the surface consists of two parallel sheets of charge, a sheet of negative electricity, of density as if there were electrons of charge 4.77×10^{-10} e.s.u., spread out uniformly with a density of one to a square 4×10^{-8} cm. on a side, and inside that at a distance of 0.5×10^{-8} cm. a similar sheet of positive charges. Remember that 300 volts = 1 e.s.u. of potential.

9. Discuss the potential and field of a dipole.

10. An uncharged metallic sphere of radius R is placed in a homogeneous electric field of intensity E_0 . Calculate the potential at any point of space, and sketch the equipotential curves. (Hint: Solve Laplace's equation in

polar coordinates taking the z axis as the direction of E_0 . Note that there is symmetry about the z axis. Try a solution of the form

$$\phi = F_1(r) + F_2(r) \cos \theta$$

with the conditions that

$$\begin{aligned} F_1(r) &\rightarrow 0, \text{ as } r \rightarrow \infty \\ F_2(r) &\rightarrow E_0 r, \text{ as } r \rightarrow \infty \end{aligned}$$

and that ϕ must be constant all over the sphere of radius R .) Solve the problem for the case that the sphere carries a total charge e .

11. The equipotentials due to two point charges e and e' are given by $e/r + e'/r' = C$. Show that the surface becomes spherical if e is of opposite sign to e' and $C = 0$. Consider a spherical conductor coinciding with this surface which is grounded. This does not disturb the field, so that these charges give the field we would have if one of the charges were removed and the metallic sphere left there. Show that if a is the radius of the sphere and L the distance from the charge (outside the sphere) to the center of the sphere, the image charge inside the sphere lies a distance L' from the center such that $a^2 = LL'$ and has a charge $e' = (-ea/L)$. Show that the surface density of induced charge varies inversely as the cube of the distance from the charge outside the surface to the point of the surface under consideration.

CHAPTER XX

MAGNETIC FIELDS, STOKES'S THEOREM, AND VECTOR POTENTIAL

The static magnetic field resembles the electrostatic field in many ways. The intensity of the field due to a magnetic pole is equal to the pole strength divided by the square of the distance of the point at which the intensity is measured, so that magnetic poles display close analogy to electric charges. The intensity of this field H is defined as the force per unit magnetic pole, and this is measured in the system of units known as the electromagnetic, as distinct from the electrostatic. We shall discuss the relation between these systems of units in a later section. The vector H satisfies the equation

$$\operatorname{div} H = 4\pi \times \text{density of magnetic poles},$$

but here a very important difference appears; north and south magnetic poles never can exist alone. No matter how small one takes a volume element, the north and south poles just cancel, so that the total density of magnetic poles is zero. Hence we have

$$\operatorname{div} H = 0. \quad (1)$$

Thus we must always deal with at least a pair of opposite poles, and here we always have a magnetic dipole, whose behavior is just like that of an electric dipole. The magnetic moment of a bar magnet is defined as the product of the strength of one of the poles times the distance of separation, and magnetic fields are measured by measuring the torque exerted on a suspended magnet (magnetometer). Exactly as we have defined the electromotive force in an electric field as $\int_A^B E \cdot ds$, we can now define as the magnetomotive force $\int_A^B H \cdot ds$. This is the work per unit pole done by the magnetic field as the pole is moved along a path from A and B . There is also a magnetic potential $\phi = -\int H \cdot ds$, and in the field of permanent magnets $\int H \cdot ds$ taken around any closed path is zero.

142. The Magnetic Field of Currents.—It is when we come to consider the magnetic fields due to currents that we meet differences from the electrostatic case. Suppose that we have a straight wire in which a steady current flows. The magnetic lines of force are concentric circles around the wire and it is clear that if we calculate the integral $\int H \cdot ds$ following one of these circles, we shall not find that its value is zero for such a closed path. On the other hand if we evaluate $\int H \cdot ds$ around any closed path which does not encircle the wire, it does vanish, and the situation is then analogous to the electrostatic case. These considerations hold for any closed circuit carrying a current. We can reduce our problem to an ordinary magnetostatic one

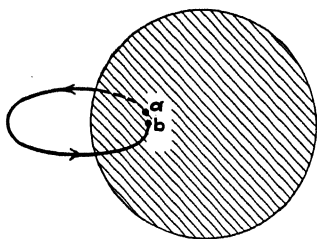


FIG. 39.—Magnetic shell and multiple valued potential. The potential difference between a and b is $4\pi m_0$, or $4\pi i$, where m_0 is the strength of the double layer producing the same magnetic field as the current i in the wire encircling the shell.

by the following device: suppose that we construct a surface bounded by the wire carrying the current and do not allow any of the curves along which we calculate $\int H \cdot ds$ to cut through this surface. Then no closed paths are possible which encircle the current, $\int H \cdot ds = 0$ around every path, and everywhere in space there is a magnetic potential Φ . Suppose we evaluate $\int H \cdot ds$ along a curve starting at a on one side of the surface and following a line of force around to a point b on the other side of the surface, as in Fig. 39. The difference of magnetic potential between a and b is given by

$$\Phi_a - \Phi_b = - \int_a^b H \cdot ds = - \int_a^b \frac{d\Phi}{ds} ds,$$

and the potential difference does not approach zero as we let a approach b since then the curve would cut our surface. This must mean that there is a jump in potential as we cross the surface. We have already seen in the last chapter that a surface distribution of dipoles (a double layer) produces a discontinuity in potential, so that we can replace our current by a surface layer of magnetic dipoles on a surface whose boundary is the current-carrying wire, and produce exactly the same magnetic field as the current. Suppose that we have a surface of area A on which we have a dipole layer of constant moment m_0

per unit area. (This may be either an electric or magnetic dipole layer). Consider a point P outside the surface. If one looks from P to the surface, the surface subtends a solid angle Ω . It is easy to show that the potential at P is equal to m_0 times Ω . The proof of this is left to a problem. In particular if P approaches the surface, Ω approaches 2π so that the potential at a point just one side of the surface is $2\pi m_0$. Similarly on the other side of the surface the potential is $-2\pi m_0$, so that there is a discontinuity of potential equal to $4\pi m_0$ as one crosses the double layer. Thus in our case we have

$$\Phi_a - \Phi_b = \mp 4\pi m_0$$

(the \pm depends on which way we go around the curve ab), so that

$$\oint H \cdot ds = \pm 4\pi m_0$$

around a closed curve which cuts through the double layer surface and is zero for every other closed curve. In the following we shall always go around the curve in such a direction that

$$\oint H \cdot ds = 4\pi m_0.$$

If we now ask how m_0 depends on the current, we must get the answer from experiment and the relation turns out to be exceedingly simple; the magnetic moment per unit area m_0 is proportional to the current. If we have not as yet defined the unit of current we may place $m_0 = i$, and this equation defines the unit current in the electromagnetic system of units. Thus

$$\oint H \cdot ds = 4\pi i \quad (2)$$

where the integration is carried once around a path encircling the wire. If we go around again the value of the integral increases again by $4\pi i$, and so on for every complete circuit of the path. This unit of current which we have introduced is called the abampere and is ten times as large as the practical unit, the ampere. On the other hand, we might wish to utilize the electrostatic unit of current, defined as the current in which one electrostatic unit of charge passes a given point per second. It is necessary to determine experimentally the proportionality constant between m_0 and i . This has been done and turns out to be $1/c$, where $c = 3 \times 10^{10}$ cm. per second. If we express our current in electrostatic measure, the work done in carrying

a unit north pole around a circuit enclosing the current is

$$\oint H \cdot ds = \frac{4\pi i}{c}. \quad (3)$$

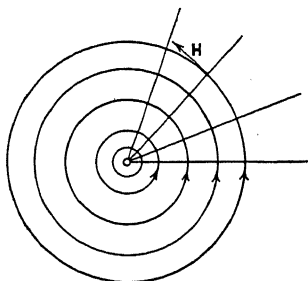
The e.s.u. of current is $\frac{1}{9} \times 10^{-9}$ ampere.

143. Field of a Straight Wire.—We can illustrate these ideas easily in the case of a straight wire carrying a constant current. Since the lines of force are circles, let us calculate the work done in carrying a unit pole around such a circle of radius r . In this case H has the same value all along the circle and is tangent to it. Thus

$$\oint H \cdot ds = H \int ds = 2\pi r H = 4\pi i$$

so that the magnetic field intensity at a distance r from a straight wire carrying a current i is

$$H = \frac{2i}{r}. \quad (4)$$



We can now set up the potential for this case.

Thus, let the wire be along the z axis, as in Fig. 40, so that the field is given by

$$H_x = \frac{-2iy}{r^2}, \quad H_y = \frac{2ix}{r^2}, \quad H_z = 0.$$

FIG. 40.—Magnetic lines of force (circles) and equipotentials (radii) for the field of a wire carrying a current (at right angles to the paper). H is perpendicular to radius. Therefore $H_x: H_y = -y:x$.

Then $\text{curl } H = 0$, as we can immediately prove by substitution.

Thus, for example, $\text{curl}_z H =$

$$\frac{\partial(2ix/r^2)}{\partial x} - \frac{\partial(-2iy/r^2)}{\partial y} = \frac{2i}{r^2} - \frac{4ix^2}{r^4} + \frac{2i}{r^2} - \frac{4iy^2}{r^4} = 0. \quad \text{Then we}$$

can have a potential, and it is easy to see that it must have as its equipotentials the lines $\theta = \text{constant}$, where θ is the polar angle in the xy plane, since these are at right angles to the lines of force. If we set $\Phi = -2i\theta = -2i \tan^{-1}(y/x)$, we have $-\partial\Phi/\partial x = -2iy/r^2 = H_x$, $-\partial\Phi/\partial y = 2ix/r^2 = H_y$, so that we have actually exhibited the potential.

But now we see that the potential is not single-valued. For a given value of x and y , the angle $\tan^{-1}(y/x)$ can have an infinite number of values, differing by 2π , and the potential can have an infinite number of values differing by $4\pi i$, in agreement with what we found before. Thus the potential is not defined in

as simple and definite a way as in electrostatics. The interpretation of this situation comes from a theorem called Stokes's theorem.

144. Stokes's Theorem.—Stokes's theorem states that if we have any closed curve, and integrate the tangential component of a vector around it, the result is equal to what we obtain if we take some surface bounded by the curve, and integrate the normal component of the curl of F over this surface:

$$\oint F_s ds = \iint \text{curl}_n F dS. \quad (5)$$

To prove it, we first divide up the surface into small surface elements, of area dS . For one of these the surface integral is $\text{curl}_n F dS$. Now suppose we choose the axes so that the z axis is normal to dS , and the area dS is bounded by $x, y, x + dx$, and $y + dy$ as in Fig. 41.

Then the surface integral is $\left(\frac{\partial F_y}{\partial x} - \right.$

$\left. \frac{\partial F_x}{\partial y} \right) dx dy$. Let us next compute

$\oint F_s ds$ for the element of area. It is evidently $F_x(x, y)dx + F_y(x + dx, y)dy$

$- F_x(x, y + dy)dx - F_y(x, y)dy = \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) dx dy$, if we go

around so as always to keep the surface on the left. Thus the theorem is true for such an infinitesimal surface. But now, if we put the whole surface together out of its elements, the total surface integral will be the sum of the parts, or $\iint \text{curl}_n F dS$. Also the total line integral will be the sum of the integrals over the separate elements. To see this, we note that in making the sum, all boundaries except the outside edge of the area are shared by two elements of the area, and the line integral from one traverses the boundary in one direction, from the other in the opposite direction, so that the contributions all cancel, leaving only the integral over the outer boundary, which is then $\oint F_s ds$. Thus Stokes's theorem is proved.

145. The Curl in Curvilinear Coordinates.—It is often useful to have the curl, and Stokes's theorem, in curvilinear coordinates. We refer back to Chap. XVIII, using methods analogous to those used there in discussing the divergence and gradient. Consider an approximately rectangular area, similar to that in Fig. 38,

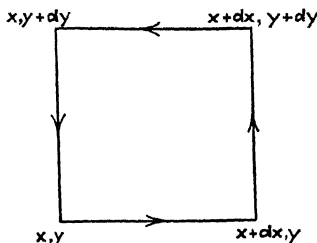


FIG. 41.—Circuit for proving Stokes's theorem.

bounded by $q_1, q_1 + dq_1, q_2, q_2 + dq_2$. The line integral about the circuit is $F_1(q_1, q_2)ds_1 + F_2(q_1 + dq_1, q_2)ds_2 - F_1(q_1, q_2 + dq_2)ds_1 - F_2(q_1, q_2)ds_2$

$$= \left[\frac{F_2(q_1 + dq_1, q_2)}{h_2} - \frac{F_2(q_1, q_2)}{h_2} \right] dq_2 - \left[\frac{F_1(q_1, q_2 + dq_2)}{h_1} - \frac{F_1(q_1, q_2)}{h_1} \right] dq_1$$

$$= \left[\frac{\partial}{\partial q_1} \left(\frac{F_2}{h_2} \right) - \frac{\partial}{\partial q_2} \left(\frac{F_1}{h_1} \right) \right] dq_1 dq_2.$$

Since this must be $\text{curl}_3 F ds_1 ds_2$, we have

$$\text{curl}_3 F = h_1 h_2 \left[\frac{\partial}{\partial q_1} \left(\frac{F_2}{h_2} \right) - \frac{\partial}{\partial q_2} \left(\frac{F_1}{h_1} \right) \right], \quad (6)$$

with analogous relations for the two other components.

We can illustrate the formulas by showing that the curl of the field of a straight wire is zero. Let us take cylindrical coordinates, in which $r = q_1, \theta = q_2, z = q_3, h_1 = 1, h_2 = 1/r, h_3 = 1$. The assumed magnetic field, along the tangent, is $H_r = 0, H_\theta = 2i/r, H_z = 0$. We then have $H_\theta/h_\theta = 2i$, a constant, so that its derivative is zero, and the curl vanishes.

146. Applications of Stokes's Theorem.—Let us apply Stokes's theorem in a few cases. First, if the curl is everywhere zero, the line integral of the vector is zero around a closed path. It follows that the line integral from one point to another along any path is the same. This is the condition for the existence of a potential, and we now see that the vanishing of the curl is just the condition that we must have in order to set up a potential. But in the magnetic case, it is not true that the line integral around any path is zero. Any contour including the current has an integral different from zero. The whole situation is then explained if inside the wire carrying the current the curl of H is not zero, but is a vector pointing along the direction of the current, of such a magnitude that the total surface integral over the cross section of the wire is $4\pi i$. Thus, for example, a contour going once around the current has a surface integral of the curl equal to $4\pi i$, which therefore must be the value of the line integral of the tangential component of H .

To find the exact relation between the current and curl H , we imagine the current in the wire to be spread out through the actual material of the wire, as in fact it is. We set up u , the

current density, or flux of electricity, satisfying the equation of continuity $\partial\rho/\partial t + \text{div } u = 0$. Then $i = \iint u_n dS$, where the integration is over the cross section of the wire. We must have, then, $4\pi \iint u_n dS = \iint \text{curl}_n H dS$, and since this must hold for any size wire, the natural assumption is that the same relation holds between the small elements of current, so that $4\pi u_n = \text{curl}_n H$, or more generally

$$\text{curl } H = 4\pi u. \quad (7)$$

Here u is in e.m.u. If it is in e.s.u., the equation is $\text{curl } H = 4\pi u/c$. We can see one result of these equations. If the current instead of being in a single wire, is distributed through space, the curl is different from zero everywhere, and there is no possibility of writing a potential at all.

147. Example: Magnetic Field in a Solenoid.—Suppose we have an infinite solenoid, of finite radius, with n turns per centimeter, carrying current i , and that we wish to calculate the magnetic field inside it. We assume that it is in no external magnetic field, so that the field outside is zero. By symmetry, the field inside will point in the direction of the axis. Now let us apply Stokes's theorem to a path as follows: (1) Inside, along a line parallel to the axis, for 1 cm. The integral of H will be H_i , the H inside, times unit distance. (2) Straight out, radially, to the outside of the solenoid. Since H is at right angles, the integral of H will be zero. (3) Outside, back for 1 cm. along a line parallel to the axis. The integral is zero since H is zero outside. (4) Straight in again, closing the figure, and contributing nothing to the integral. Thus we have $\oint H_s ds = H_i$. Now $\iint \text{curl}_n H dS = 4\pi \iint u_n dS = 4\pi \times \text{total current flowing through the contour} = 4\pi ni$. Hence we have $H_i = 4\pi ni$, the formula for the magnetic field inside a solenoid, showing that it is constant independent of position.

148. The Vector Potential.—In magnetic fields coming from permanent magnets, where there is no current, we can write an ordinary potential letting $H = -\text{grad } \Phi$. But this is only possible when $\text{curl } H = 0$, which is not true in the presence of currents. On the other hand, it can be shown that if the divergence of a vector is zero, as $\text{div } H = 0$, it is always possible to set up a vector A , called the vector potential (to distinguish it from Φ , which is called a scalar potential), such that $H = \text{curl } A$. This is often a useful thing to do. We can prove readily that $\text{div curl } A = 0$ always, so that we have $\text{div } H = 0$.

The vector potential satisfies a simple differential equation. We know that $\text{curl } A = H$, but this does not determine A uniquely. In fact, to determine a vector field uniquely we must specify both its curl and its divergence, and we can find a vector whose curl and divergence are any desired functions. Let us then demand that $\text{div } A = 0$. We now have $\text{curl } H = 4\pi u/c = \text{curl curl } A$. It can be proved that $\text{curl curl } A = \text{grad div } A - \nabla^2 A = -\nabla^2 A$, since $\text{div } A = 0$. Hence

$$\nabla^2 A = -\frac{4\pi u}{c}, \quad (8)$$

similar to Poisson's equation for the scalar potential in terms of the charge density,

$$\nabla^2 \phi = -4\pi\rho.$$

These two equations, expanded to include terms depending on time, prove to be very important in general electrical theory.

Let us set up the vector potential for a current in a straight wire. Take cylindrical coordinates, with the wire pointing along the z axis. Poisson's equation for A is a vector equation, but since u has only a z component, A will likewise have only a z component, which will depend only on r . Thus we have

$$\begin{aligned} \frac{1}{r} \frac{d}{dr} \left(r \frac{dA_z}{dr} \right) &= -\frac{4\pi u}{c} \text{ for } r < R, \\ &= 0 \text{ for } r > R. \end{aligned}$$

where R is the radius of the wire.

The solutions of this equation are

$$\begin{aligned} A_z &= -\frac{\pi r^2 u}{c} + a \ln r + b \text{ for } r < R \\ &= d \ln r + e \quad \text{for } r > R. \end{aligned}$$

Since A cannot become infinite at $r = 0$, we must have $a = 0$. We may choose $b = 0$. Then d and e must be chosen to make A and its derivative with respect to r continuous at $r = R$. Noting that $\pi R^2 u = i$, the total current, this easily leads to

$$A_z = -\frac{2i}{c} \ln r + \text{constant} \quad \text{for } r > R.$$

The only component of H is then H_θ , which is

$$H_\theta = h_r h_z \left[\frac{\partial}{\partial z} \left(\frac{A_r}{h_r} \right) - \frac{\partial}{\partial r} \left(\frac{A_z}{h_z} \right) \right] = \frac{2i}{cr}.$$

149. The Biot-Savart Law.—In the case of a linear conductor carrying a current i , the expression for the vector potential, using the solution of Poisson's equation from Chap. XIX, becomes

$$A = \frac{i}{c} \int \frac{ds}{r},$$

where ds is the vector element of length taken along the conductor, and pointing in the direction of the current. To find the intensity of the magnetic field, we take the curl, finding

$$H = \text{curl } A = \frac{i}{c} \int \text{curl } \frac{ds}{r}.$$

In this equation, ds is a vector and r a scalar. In general, if S is a scalar and B an arbitrary vector, it is easy to show that

$$\text{curl } (SB) = S \text{ curl } B + (\text{grad } S) \times B.$$

Applying this relation to our case, $B = ds$, and $S = 1/r$, and we must remember that in taking the curl we differentiate only with respect to the coordinates which fix the point at which we wish the value of H (the field point). Now these coordinates appear only in r and not in ds , which depends on the circuit only. Thus the first term vanishes and we have

$$H = \frac{i}{c} \int \text{grad } \left(\frac{1}{r} \right) \times ds = \frac{i}{c} \int \frac{ds \times \vec{r}}{r^3} \quad (9)$$

where \vec{r} is the vector from ds to the field point, and r is the length of this vector. If we imagine that the resultant H is made up of a sum of contributions from each conductor element ds , we may write the law in its differential form

$$dH = \frac{i}{cr^3} (ds \times \vec{r}). \quad (10)$$

This is known as the Biot-Savart law. The magnitude of dH is obviously

$$|dH| = \frac{i}{cr^2} ds \sin \theta, \quad (11)$$

where θ is the angle between the direction of ds and r ; the direction of dH is perpendicular to the plane of ds and r . Applied to closed circuits it always yields the same results as the integral law. For open circuits this is not obvious, since we can add to the expressions for dH a differential $d\psi$ provided $\oint d\psi$ around a

closed curve is zero. In this way we leave the law for closed circuits unaltered, but for open circuits change the value of H so calculated. Thus the integral law must be looked upon as the more fundamental.

Problems

1. Prove that a double layer of moment m_0 per unit area leads to a potential ϕ at point P equal to $m_0\Omega$, where Ω is the solid angle subtended by the area from the point P .

2. Show that in the electrostatic system of units, charge has the dimension $m^{1/2}l^{3/2}t^{-1}$, current the dimensions $m^{1/2}l^{3/2}t^{-2}$, voltage (e.m.f.) the dimensions $m^{1/2}l^{1/2}t^{-1}$, resistance the dimensions $l^{-1}t$, and capacity the dimensions l .

3. Derive the dimensions of charge, current, voltage, resistance, and capacity in the electromagnetic system of units.

4. Prove that if S is a scalar and B a vector

$$\text{curl}(SB) = S \text{curl} B + \text{grad} S \times B.$$

5. Prove $\text{div} \text{curl} F = 0$; $\text{curl} \text{curl} F = \text{grad} \text{div} F - \nabla^2 F$, where F is any vector.

6. Using the Biot-Savart law, find the magnetic field at any point on the axis of symmetry of a circular loop of wire of radius R carrying a current i .

7. A current flows in a circular loop of wire, of radius R . Find the vector potential of the resulting magnetic field, at large distances compared with R , by adding the contributions to the vector potential due to the separate elements of current.

8. Compute the field, from the potential of the last problem, and show that it is approximately the field of a single dipole. Find the strength of the dipole, in terms of current and radius R .

9. Two parallel straight wires carry equal currents. Work out the magnetic fields due to the two together, in the two cases where the currents flow in the same or in opposite directions, drawing diagrams of the lines of force.

10. Find the magnetic field at points inside a wire carrying a current, assuming the wire is straight and of circular cross section and that the current has constant density throughout the wire.

11. Compute the curl in spherical polar coordinates. Verify directly that the divergence of a curl is zero in these coordinates.

CHAPTER XXI

ELECTROMAGNETIC INDUCTION AND MAXWELL'S EQUATIONS

We now leave the restriction of the steady state and inquire into the extensions of the theory necessary to have it hold for nonstationary phenomena. The fundamental fact concerning electromagnetic induction may be stated as follows: If a set of circuits carrying current (or magnets and circuits) are set in relative motion with respect to each other, the currents in the circuits change during the relative motion. Instead of formulating a law for the induced currents, it is simpler to consider the induced electromotive force. Take a closed circuit in the neighborhood of a moving magnet (or moving circuit), and let N be the number of magnetic lines of force through the circuit. Then the induced electromotive force is $-\frac{dN}{dt}$, expressed in electromagnetic units, if N is in these units. If the e.m.f. is expressed in electrostatic units it is equal to $-\frac{1}{c} \frac{dN}{dt}$. The minus sign expresses what is commonly termed Lenz's law and indicates that if $\frac{dN}{dt}$ is represented by a vector going through the circuit, the induced current flows in a clockwise fashion.

150. The Differential Equation for Electromagnetic Induction.

We can now state this law in more analytical form. Consider the closed curve formed by the circuit, and any surface whose boundary is this curve, so that the surface forms a sort of cap over the curve. Then the magnetic flux

$$N = \iint H_n dS$$

where the integral is carried out over the whole surface. Furthermore the electromotive force is by definition the work done in carrying a unit charge once around the circuit. This work may be done either by the electric field or by chemical forces in a battery. Since the latter are considered absent we have

$$\text{e.m.f.} = \oint E_s ds$$

where the integral is taken completely around the circuit. The fact that this line integral does not vanish shows us at once that we shall not be able to introduce a potential, as we have done in the electrostatic case. Thus we have

$$\oint E_s ds = -\frac{d}{dt} \iint H_n dS. \quad (1)$$

It should be noticed that the flux of the magnetic field through the circuit may change in several ways, either by changing H_n , or by changing the shape of the circuit, thus causing a change in the enclosed area, or by moving the undeformed circuit to other parts of space where H_n is different. In general dN/dt is composed of several terms. In the case of fixed circuits, we may replace the total time derivative by the partial derivative so that $dN/dt = \partial N/\partial t$. With the help of Stokes's theorem we rewrite the induction law as

$$\iint \text{curl}_n E dS = -\frac{d}{dt} \iint H_n dS = -\iint \frac{\partial H_n}{\partial t} dS.$$

This holds for any fixed circuit, and hence for any fixed area of integration. Thus it must hold for an infinitesimal area dS , so that the integrands must be equal and we obtain

$$\text{curl } E = -\frac{\partial H}{\partial t}.$$

This is the differential form of the induction law. In it, E and H are both expressed in e.m.u. If E is expressed in e.s.u. and H in e.m.u., the law takes the form

$$\text{curl } E = -\frac{1}{c} \frac{\partial H}{\partial t}. \quad (2)$$

151. The Displacement Current.—We have now derived four fundamental electromagnetic equations:

$$\begin{aligned} \text{div } E &= 4\pi\rho, \\ \text{div } H &= 0, \\ \text{curl } E &= -\frac{1}{c} \frac{\partial H}{\partial t}, \\ \text{curl } H &= \frac{4\pi u}{c}, \end{aligned} \quad (3)$$

where E , ρ , and u are in e.s.u. and H in e.m.u. These are almost the Maxwell equations, but there is difficulty with the last of

them. Of course, we have derived it on the basis of steady closed currents and for this case it is surely correct. The difficulty occurs when we try to apply this result to nonstationary cases. In the nonsteady state we have the new possibility of current flowing in "open" circuits. The simplest example is that of the discharge of a condenser. Here the current starts at the positively charged plate, whose charge diminishes as the current flows to the negatively charged plate and annuls the charge there. Thus we can look upon the condenser plate as a source (or sink) of current. Now if we take the divergence of the last equation, we have

$$\operatorname{div} \operatorname{curl} H = \frac{4\pi}{c} \operatorname{div} u$$

and since the divergence of any curl is zero, we find that $\operatorname{div} u$ equals zero, which means that the current is always closed and there are no sources or sinks. Thus open circuits lead to a contradiction to this equation. We have derived the equation from steady-state considerations, however, and if we are to extend it to hold under all conditions, it is clear that there must be some term which vanishes for the steady state which we must add. The equation of continuity applied to electric charge and current tells us that

$$\operatorname{div} u + \frac{\partial \rho}{\partial t} = 0$$

expressing the fact that the flow of current out of a volume results in a decrease of charge in that volume. In the steady state $\partial \rho / \partial t = 0$, so that $\operatorname{div} u = 0$, and we have no inconsistency with our fundamental equation. It is certainly clear that if $\operatorname{curl} H$ is to be proportional to a current, this current must be divergenceless, and u is not. Maxwell made the bold step of assuming that the whole current consisted of two terms u and u' , where u' was so chosen that $\operatorname{div} (u + u') = 0$. In this way the distinction between open and closed circuits vanishes and a unity hitherto lacking was given to the laws. Maxwell saw at once

that we must set $u' = \frac{1}{4\pi} \frac{\partial E}{\partial t}$. For then we have

$$\begin{aligned} \operatorname{div} (u + u') &= \operatorname{div} \left(u + \frac{1}{4\pi} \frac{\partial E}{\partial t} \right) = \operatorname{div} u + \frac{1}{4\pi} \frac{\partial}{\partial t} (\operatorname{div} E) \\ &= \operatorname{div} u + \frac{\partial \rho}{\partial t} = 0 \end{aligned}$$

and this is the equation of continuity which we have been trying to satisfy. In other words, Maxwell assumed the correct equation to be

$$\text{curl } H = \frac{1}{c} \frac{\partial E}{\partial t} + \frac{4\pi}{c} u. \quad (4)$$

The new term $\frac{1}{4\pi} \frac{\partial E}{\partial t}$ is called the displacement current, in contrast to the convection current u .

Actually the real advance of Maxwell over his predecessors lies in the introduction of this displacement current. The physical meaning of this current can be obtained by considering the charging of a condenser. Current flows from one plate through the wire to the other plate. If the current is i , this equals the rate of increase of charge on the plate. Suppose the plates are of area A , separation d , then the field between them is

$$E = 4\pi\sigma = \frac{4\pi}{A} \times \text{total charge}$$

and the displacement current density in the region between the plates is

$$\frac{1}{4\pi} \frac{\partial E}{\partial t} = \frac{\partial \sigma}{\partial t} = \frac{1}{A} \frac{\partial}{\partial t} (\text{total charge}) = \frac{i}{A}.$$

Thus the displacement current is $\frac{A}{4\pi} \frac{\partial E}{\partial t} = i$, and is equal to the convection current in the wire, so that the current becomes continuous throughout the circuit. The fundamental assumption of Maxwell was that the displacement current is always present when an electric field varies in time and produces the same magnetic effects as convection currents.

It is clear that a test of Maxwell's hypothesis can only be made with very rapidly varying fields, since we must make $\frac{1}{4\pi} \frac{\partial E}{\partial t} \gg u$ in order to keep the convection current effects from masking the displacement current effects. As is well known, Hertz, in 1888, performed the experiments on electric waves which confirmed this assumption of Maxwell. There is an interesting connection between the displacement current and the Biot-Savart law. All the attempts before Maxwell were to find a correct form of the Biot-Savart law for "open" circuits. As we pointed out in the last chapter, the addition of a total differential to this law would yield nothing when it was applied to closed

circuits, and the hope was that the correct form to be added to this law could be found so as to account for open circuit phenomena.

152. Maxwell's Equations.—We can now write the correct Maxwell equations

$$\begin{aligned}\text{curl } H &= \frac{1}{c} \frac{\partial E}{\partial t} + \frac{4\pi u}{c} & \text{curl } E &= -\frac{1}{c} \frac{\partial H}{\partial t} \\ \text{div } H &= 0 & \text{div } E &= 4\pi\rho.\end{aligned}$$

These are the fundamental equations of electromagnetic theory. They need extension in but one way. If there are dielectric and magnetic bodies present, in them Coulomb's law and its analogue for the magnetic field become

$$F = \frac{ee'}{\epsilon r^2},$$

and

$$F = \frac{mm'}{\mu r^2},$$

where ϵ is the dielectric constant and μ the magnetic permeability. We now introduce a new vector called the electric displacement D , defined by $D = \epsilon E$, where E is the intensity of the electric field. Similarly, we introduce the magnetic induction vector $B = \mu H$. It is easy to see from our previous work that we now have the relation $\text{div } D = 4\pi\rho$. Furthermore, Faraday's induction law refers to the rate of change of magnetic flux through a circuit and hence H must be replaced by B in this relation.

Finally, we have $\text{div curl } E = 0 = -\frac{1}{c} \frac{\partial}{\partial t} \text{div } B$, so that $\text{div } B = 0$, rather than $\text{div } H = 0$. The final equations are thus found to be:

$$\begin{aligned}\text{curl } H &= \frac{1}{c} \frac{\partial D}{\partial t} + \frac{4\pi u}{c} & \text{curl } E &= -\frac{1}{c} \frac{\partial B}{\partial t} \\ \text{div } B &= 0 & \text{div } D &= 4\pi\rho \\ B &= \mu H & D &= \epsilon E.\end{aligned}\tag{5}$$

In these equations, E , D , ρ , u are in electrostatic units, H and B in electromagnetic units. In Chap. XXIV we discuss in detail the significance of B and D , and the interpretation of ϵ and μ .

Maxwell's equations suffice to determine the field, when we are given the charges and currents. To make a complete set of dynamical principles, however, we need two more relations.

First is the formula giving the force acting on a charge and current. The electrical force per unit volume is simply ρE , the force on unit charge multiplied by the charge per unit volume. The magnetic force is that acting on the current, as observed in the ordinary action of the electric motor. This force acts at right angles both to the current and to the magnetic field, and is equal, as is shown in the elementary study of electricity, to the current (in electromagnetic units) times the component of magnetic field at right angles to the current. For unit volume, this is just given by the vector product $u \times H$. If u is in electrostatic units, it is $\frac{u}{c} \times H$. Thus we have for the force vector

$$F = \rho E + \frac{1}{c}(u \times H).$$

If the current density is produced by the motion of charge, we have $u = \rho v$, where v is the velocity vector of the charge. In this case

$$F = \rho \left[E + \frac{1}{c}(v \times H) \right].$$

This relation has been particularly used by Lorentz.

Finally, one must have a law, such as Newton's law stating that the force is equal to mass times acceleration, determining the motion of charge in terms of the force acting. With such a law, we find the field from the charge, the force from the field, and the motion from the force, obtaining therefore a complete system of dynamics.

Let us now summarize the various steps gone through in building up Maxwell's equations. Consider first the static case. Here $\partial D/\partial t = \partial B/\partial t = 0$ and $u = 0$. The equations become

$$\begin{array}{ll} \text{curl } H = 0 & \text{curl } E = 0 \\ \text{div } B = 0 & \text{div } D = 4\pi\rho \\ B = \mu H & D = \epsilon E. \end{array}$$

The three equations on the left are those of magnetostatics, and the remaining three are those of electrostatics. Each system is completely independent of the other. The equations $\text{curl } H = 0$, and $\text{curl } E = 0$, show that scalar potentials exist.

In the stationary case, we still have $\partial B/\partial t = \partial D/\partial t = 0$, but now $u \neq 0$. The only one of the equations above which is modified is $\text{curl } H = 4\pi u/c$, the others remaining unchanged.

It is usual to include Ohm's law in the statement of the equations, however. This law is easily stated in differential form by considering a small volume, having length L in the direction of the current flow, and cross-sectional area A normal to the current. We apply Ohm's law in the form $p.d. = iR$. Here the potential difference is the field E times the length L of the volume, the current is the area times the current density u , and the resistance is the specific resistance times L/A . Hence we have

$$EL = Au \frac{L}{A} \times \text{specific resistance},$$

or

$$u = \sigma E, \quad (6)$$

where σ , the specific conductivity, is the reciprocal of the specific resistance. This equation is Ohm's law in the form suitable for Maxwell's equations, and it is commonly included along with $D = \epsilon E$ and $B = \mu H$.

If we now proceed to the nonstationary state we must strictly use the correct Maxwell relations. But there is a case of utmost practical importance, in which $\partial D / \partial t \ll 4\pi u$, and hence for which the effects of displacement can be neglected in comparison with those of the convection currents. The Maxwell equations with the displacement current omitted apply to the so-called "quasi-stationary" processes, and these form practically the whole domain of electrical engineering. The magnetic field inside and outside conductors is calculated as if produced only by the convection currents, but the induction law is not left out as in the stationary state. Here we have a double coupling of electric and magnetic fields, first, as in the stationary case, where electric currents produce magnetic fields, and, secondly, by the induction law. Since the essentially new contribution of Maxwell, the displacement current, is neglected in quasi-stationary calculations, it is clear that no study in that field can give experimental confirmation of Maxwell's idea.

153. The Vector and Scalar Potentials.—We observe that, if H depends on time, $\text{curl } E \neq 0$, so that there is no potential for E . The ordinary electrical potential is thus confined to static problems. Further, if u or $\frac{\partial E}{\partial t} \neq 0$, there is no potential for H .

We have seen in the last chapter how a potential can be intro-

duced for H : one uses a vector potential A , possible because $\text{div } H = 0$. That is, we let

$$H = \text{curl } A. \quad (7)$$

We can do this even in the general case. And it proves that we can use a scalar potential ϕ , reducing to the electrostatic potential in the case of a steady state, but different in other cases, by a special device. The relation which proves to be satisfied is that

$$E = -\text{grad } \phi - \frac{1}{c} \frac{\partial A}{\partial t}, \quad (8)$$

reducing to the familiar $E = -\text{grad } \phi$ when everything is independent of time. These relations are written for the case of empty space, where $\epsilon = \mu = 1$, and we shall give the discussion only for that case.

To verify our statements about the vector potential A and the scalar potential ϕ we substitute the expressions for E and H in Maxwell's equations, and see if they can be satisfied by the proper choice of A and ϕ . First, we notice that $\text{div } H = \text{div } \text{curl } A = 0$, so that this equation is automatically satisfied.

Next we take $\text{div } E = -\text{div } \text{grad } \phi - \frac{1}{c} \frac{\partial}{\partial t} \text{div } A = -\nabla^2 \phi - \frac{1}{c} \frac{\partial}{\partial t} \text{div } A$. This must equal $4\pi\rho$. Now we consider the curl

equations. We have $\text{curl } E = -\text{curl } \text{grad } \phi - \frac{1}{c} \frac{\partial}{\partial t} \text{curl } A$.

Since the curl of any gradient is zero, this is $-\frac{1}{c} \frac{\partial}{\partial t} \text{curl } A =$

$-\frac{1}{c} \frac{\partial H}{\partial t}$, verifying another of Maxwell's equations. Finally

$\text{curl } H = \text{curl } \text{curl } A = \text{grad } \text{div } A - \nabla^2 A$. This must equal $\frac{1}{c} \frac{\partial E}{\partial t} + \frac{4\pi u}{c} = -\frac{1}{c} \frac{\partial}{\partial t} \text{grad } \phi - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} + \frac{4\pi u}{c}$. Hence, in order to satisfy Maxwell's equations, we must have

$$-\nabla^2 \phi - \frac{1}{c} \frac{\partial}{\partial t} \text{div } A = 4\pi\rho,$$

$$\text{grad } \text{div } A - \nabla^2 A + \frac{1}{c} \frac{\partial}{\partial t} \text{grad } \phi + \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} = \frac{4\pi u}{c}.$$

But now let us choose A and ϕ subject to the condition that

$\text{div } A + \frac{1}{c} \frac{\partial \phi}{\partial t} = 0$. Since $\text{div } A$ is so far arbitrary, we can do this. Then the first equation becomes

$$\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} = -4\pi\rho,$$

and the second

$$\nabla^2 A - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} = \frac{-4\pi u}{c}. \quad (9)$$

These are the equations for the potentials. If A and ϕ satisfy them, then, as we stated before, the fields determined from them by the equations $E = -\text{grad } \phi - \frac{1}{c} \frac{\partial A}{\partial t}$, $H = \text{curl } A$, satisfy Maxwell's equations. The equations for the potentials are of the form called D'Alembert's equation, and as can be seen are extensions of Poisson's equation, obtained by adding the time derivatives. We observe that in regions where there is no charge and current density, the potential satisfies the wave equation, which is the homogeneous equation obtained by setting the right side of D'Alembert's equation equal to zero. That is, ϕ and A are given by functions representing waves traveling with velocity c . Hence the same thing must be true of the fields E and H . This is the origin of the theory of electromagnetic waves, and of the electromagnetic theory of light, and the proof that c , the ratio of the units, is at the same time the velocity of light.

In regard to our condition imposed on the potentials, that $\text{div } A + \frac{1}{c} \frac{\partial \phi}{\partial t} = 0$, we can readily show that if the potentials satisfy Eqs. (9) above, this condition can also be satisfied. For take $1/c$ times the time derivative of the first, and the divergence of the second, and add. Using the fact that $\text{div } \nabla^2 A = \nabla^2 \text{div } A$, where A is any vector, the result is

$$\begin{aligned} \nabla^2 \left(\text{div } A + \frac{1}{c} \frac{\partial \phi}{\partial t} \right) - \frac{1}{c^2} \frac{\partial^2 (\text{div } A + 1/c \partial \phi / \partial t)}{\partial t^2} \\ = \frac{-4\pi}{c} \left(\text{div } u + \frac{\partial \rho}{\partial t} \right) = 0. \end{aligned}$$

That is, the quantity $\text{div } A + \frac{1}{c} \frac{\partial \phi}{\partial t}$ satisfies the wave equation everywhere. It can be proved that no function, other than zero, can satisfy the wave equation everywhere, unless its value

at infinity is different from zero. Hence in an ordinary problem of charges at finite points, where certainly the potentials must vanish at infinity, it must be that $\text{div } A + \frac{1}{c} \frac{\partial \phi}{\partial t} = 0$, and in other cases we can certainly choose the potentials so that this condition will be satisfied.

Problems

1. Show that E and H satisfy the wave equations $\nabla^2 E - \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2} = 0$, with a similar equation for H , in empty space, where u and ρ are zero, and $\epsilon = \mu = 1$. (Suggestion: for the first, take the equation for $\text{curl } E$, and take its curl, then substitute for $\text{curl } H$ in terms of E . Proceed in an analogous way with the other equation.)

2. In a region where u and ρ are zero, but ϵ and μ are different from 1, show that the velocity of light is $\frac{c}{\sqrt{\epsilon\mu}}$.

3. A magnetic field points along the z axis, and its magnitude is proportional to the time, and the same at all points of space. Find the vector potential. Assuming that the scalar potential is zero, find the induced electric field. Prove by direct integration using a circular circuit, that the law of induction holds.

4. Describe the magnetic field between the plates of a condenser while it is charging up.

5. Starting from the induction law, show that the line integral of $(E + \frac{1}{c} \frac{\partial A}{\partial t})$ around a closed path is zero, where A is the vector potential. From this show that the curl of the above vector vanishes and hence that $E = -\text{grad } \phi - \frac{1}{c} \frac{\partial A}{\partial t}$, where ϕ is the scalar potential.

6. In conductors where $\mu = 1$ and $\rho = 0$ show that E and H both satisfy differential equations of the form

$$\nabla^2 E - \frac{4\pi\sigma}{c^2} \frac{\partial E}{\partial t} - \frac{\epsilon}{c^2} \frac{\partial^2 E}{\partial t^2} = 0.$$

7. Derive the differential equations satisfied by E and H for quasi-stationary processes.

8. Show that if a voltage is induced in a circuit (2) by a changing magnetic field due to a circuit (1), the induced e.m.f. in (2) is given by

$$\oint E_{2s} ds_2 = -\frac{1}{c} \frac{d}{dt} \oint A_1 \cdot ds_2 = -\frac{\mu}{c} \frac{d}{dt} \iint H_{n1} dS_2,$$

where A_1 is the vector potential at the element ds_2 due to the current in circuit (1). For quasi-stationary processes we can write

$$A_1 = \frac{\mu}{c} \iint \int \frac{u_1 dv_1}{r_{12}},$$

where u_1 is the current density in circuit (1) and dv_1 a volume element thereof. For linear currents show that the induced e.m.f. is then given by

$$\oint E_2 ds_2 = -\frac{\mu}{c^2} \frac{d}{dt} \left(I_1 \oint \oint \frac{ds_1 \cdot ds_2}{r_{12}} \right),$$

where I_1 is the current in the first circuit, r_{12} is the distance between ds_1 and ds_2 .

The coefficient of mutual induction M_{12} is defined as

$$\mu \iint \frac{ds_1 \cdot ds_2}{r_{12}}$$

so that the above relation becomes

$$(E.m.f.)_2 = -\frac{1}{c^2} \frac{d}{dt} (M_{12} I_1).$$

9. Apply the reasoning of Prob. 8 to a single circuit and show that the self-induced voltage is

$$E.m.f. = -\frac{1}{c^2} \frac{d}{dt} (L_1 I_1)$$

where $L_1 = \mu \iint \frac{ds_1 \cdot ds_1'}{r_{11}}$, ds_1 and ds_1' being two elements of length of the conductor.

CHAPTER XXII

ENERGY IN THE ELECTROMAGNETIC FIELD

The idea of energy is as useful in electromagnetic theory as in mechanics. Maxwell's equations correspond in a general way to the equations of motion, and in the present chapter we introduce electrical and magnetic energies analogous to the potential and kinetic energies. The analogy is particularly close with the mechanical energy in a vibrating medium, since electrical oscillations in free space, as in a light wave, are similar to mechanical oscillations in sound. The energy of an elastic solid is distributed throughout the body, each volume element having a potential energy on account of its strain, and a kinetic energy on account of its velocity. Correspondingly we shall find that the electromagnetic energy can be considered as localized throughout the field, with a definite density of electrical and magnetic energy. Finally, the potential energy is proportional to the square of the stress or strain, and kinetic energy proportional to the square of velocity or momentum, and in a similar way here we shall find electrical energy proportional to the square of E or D , and the magnetic energy to the square of H or B . The analogy can be carried out completely, Maxwell's equations, for instance, being written in the form of Lagrangian equations; however, we shall not do this. We start the discussion by deriving the electrical and magnetic energy by elementary means from the condenser and solenoid, and then pass to general theorems involving energy density and energy flow.

154. Energy in a Condenser.—Given a condenser of capacity C , let its charge at a given moment be q . Assume that we are charging up the condenser, and that we wish to know how much work we shall have to do on it to charge it. To take a small additional charge dq around the circuit, against the difference of potential q/C , will require an amount of work $(q/C)dq$. Thus the whole work done in setting up a charge Q is

$$\int_0^Q \frac{q}{C} dq = \frac{1}{2} \frac{Q^2}{C}. \quad (1)$$

This is the expression for the energy in a condenser which we found in Chap. V, Prob. 6.

But now there is an interesting way in which we may consider this. We may imagine that the energy resides directly in the electromagnetic field, between the condenser plates. Let the area of the plates be A , the distance of separation d , and the dielectric constant ϵ , so that $C = A\epsilon/4\pi d$. Also the field between the plates will be $E = q/Cd$, the difference of potential between the plates divided by the distance. Hence we have $\frac{1}{2}q^2/C = \frac{1}{2}E^2Cd^2 = (\epsilon E^2/8\pi)(Ad)$. But Ad is simply the volume of the condenser, or of the region of space where the field is E . Hence we may consider the energy to be located in the electromagnetic field, with a volume density $\epsilon E^2/8\pi$, and the integral of this over the condenser will give precisely the total energy.

155. Energy in the Electric Field.—It is not difficult to show that in an arbitrary electrostatic field the energy is given by

$$\frac{\epsilon}{8\pi} \iiint E^2 dv.$$

Let us consider two point charges e_1 and e_2 in a medium of dielectric constant ϵ separated by a distance r_{12} . The force acting on each is given by Coulomb's law as

$$F = \frac{e_1 e_2}{\epsilon r_{12}^2}$$

and the potential energy of the system by

$$V = \frac{e_1 e_2}{\epsilon r_{12}} = \frac{1}{2} \left(e_1 \cdot \frac{e_2}{\epsilon r_{12}} + e_2 \cdot \frac{e_1}{\epsilon r_{12}} \right).$$

We have written this in two terms and notice that the first is just the charge e_1 times the potential at the point where the charge is due to the charge e_2 . Similarly the second term is e_2 times the potential at e_2 due to e_1 . Thus we can write

$$V = \frac{1}{2}(e_1 \varphi_1 + e_2 \varphi_2)$$

where φ_1 and φ_2 are the potentials. In general for n charges we have

$$V = \frac{1}{2} \sum_k e_k \varphi_k \quad (2)$$

and if the charges are distributed in space instead of being point charges, this becomes an integral

$$V = \frac{1}{2} \iiint \rho \varphi dv \quad (3)$$

where ρ is the density of charge. Now, by Poisson's equation we know that $\nabla^2\phi = -4\pi\rho/\epsilon$, so that the integral can be written

$$V = \frac{-\epsilon}{8\pi} \iiint \phi \nabla^2 \phi \, dv.$$

We now make use of Green's theorem in its first form

$$\iiint \psi \nabla^2 \phi \, dv + \iiint \text{grad } \psi \cdot \text{grad } \phi \, dv = \iint \psi \text{grad}_n \phi \, dS$$

where ϕ and ψ are any two scalar quantities. Place $\psi = \phi = \varphi$ and this becomes

$$\iiint \varphi \nabla^2 \varphi \, dv + \iiint E^2 \, dv = \iint \varphi \text{grad}_n \varphi \, dS$$

since $E = -\text{grad } \varphi$.

Now since we integrate over all space, we must examine the behavior of the surface integral as the surface (a sphere of radius R , for example) gets larger and larger. The potential φ varies as $1/R$ for large R , $\text{grad}_n \varphi$ as $1/R^2$ and dS is proportional to R^2 , so the whole surface integral vanishes as $R \rightarrow \infty$. Thus substituting in our expression for V , we find

$$V = \frac{\epsilon}{8\pi} \iiint E^2 \, dv \quad (4)$$

which is the equation we set out to derive. From our derivation it is easy to show that if ϵ is not constant

$$V = \frac{1}{8\pi} \iiint E \cdot D \, dv$$

where D is the electric displacement vector. This shows us the origin of the name for D . If we think of D as an ordinary displacement (per unit volume) of electricity, then the work done per unit volume is the scalar product of the force times the displacement. In an infinitesimal displacement dD , the work per unit volume is proportional to

$$E \cdot dD = \epsilon E \cdot dE$$

and for a finite displacement D we get something proportional to

$$\int_0^D E \cdot dD = \int_0^E \epsilon E \, dE = \frac{\epsilon E^2}{2} = \frac{E \cdot D}{2}.$$

Thus, except for the numerical factor $1/4\pi$, we have the potential energy per unit volume.

156. Energy in a Solenoid.—In a similar way, we may consider the magnetic energy in a solenoid to reside in the magnetic

field within the coil. We have found earlier that the energy in a solenoid of self-induction L , in which a current i was flowing, was $\frac{1}{2}Li^2$. But now we can easily write this in terms of the field H within the solenoid. We have seen that this field is $4\pi ni$, where n is the number of turns of the coil per centimeter. The coefficient of self-induction L for a coil is easily found. By definition, it is the e.m.f. induced when there is unit time rate of change of current through the coil. The e.m.f. per turn $= \frac{d}{dt} (B \times \text{cross-sectional area}) = \pi r^2 \mu \frac{dH}{dt}$, if r is the radius of the coil, μ the permeability. Thus the e.m.f. for the whole N turns is $N\pi r^2 \mu \frac{dH}{dt}$. Since $H = 4\pi ni = \frac{4\pi N i}{d}$, if N is the whole number of turns, d the length, the e.m.f. is $N\pi r^2 \mu \frac{(4\pi N)}{d} \frac{di}{dt}$, so that $L = \frac{4\pi^2 N^2 \mu r^2}{d}$. Hence we have $\frac{1}{2}Li^2 = \frac{(2\pi^2 N^2 \mu r^2)}{d} \left(\frac{Hd}{4\pi N} \right)^2 = \frac{\mu H^2 r^2 d}{8} = \frac{\mu H^2 (\pi r^2 d)}{8\pi}$. Since $\pi r^2 d$ is the volume, this indicates a volume density of magnetic energy of $\frac{H^2 \mu}{8\pi}$.

The proof that the total magnetic energy in a magnetostatic field is $\frac{\mu}{8\pi} \iiint H^2 dv$ or $\frac{1}{8\pi} \iiint H \cdot B \, dv$ is carried out in exactly the same manner as the one for the electrostatic energy given in the last paragraph.

157. Energy Density and Energy Flow.—The examples we have considered suggest that in a combined electric and magnetic field there should be a volume density $(1/8\pi)(\epsilon E^2 + \mu H^2)$ of electromagnetic energy. As a matter of fact, it proves to be quite possible to make this assumption, and to carry it out in a logical way. One can regard the electromagnetic energy almost as a fluid, having a certain density, flowing from place to place in the field. Thus, there is a flow vector associated with it, called Poynting's vector, which we shall show in the next section to be equal to $(c/4\pi)(E \times H)$. We shall prove that there is an equation of continuity for the energy:

$$\text{div} \left[\frac{c}{4\pi} (E \times H) \right] + \frac{\partial}{\partial t} \left[\frac{1}{8\pi} (\epsilon E^2 + \mu H^2) \right] = 0. \quad (5)$$

This is only true, however, in regions where electromagnetic energy is not being produced. Of course, energy as a whole is

conserved, but there can easily be sources and sinks of electromagnetic energy. Thus batteries are sources, in which chemical energy is converted into electrical energy, and resistances are sinks, in which the electrical energy is converted into heat. We imagine the field as being worked on by the battery, and as doing work against the frictional resistance. Hence our whole relation is that $\partial/\partial t$ (electromagnetic energy) = rate of production of energy from e.m.f. per unit volume—rate of dissipation of energy into heat—div (energy flow). This equation, put in mathematical form, is Poynting's theorem.

158. Poynting's Theorem.—Let us compute the quantity

$$\operatorname{div} \left[\frac{c}{4\pi} (E \times H) \right] + \frac{\partial}{\partial t} \left[\frac{1}{8\pi} (\epsilon E^2 + \mu H^2) \right].$$

It can be shown in general that

$$\operatorname{div} (A \times B) = B \cdot \operatorname{curl} A - A \cdot \operatorname{curl} B.$$

Also $\frac{\partial A^2}{\partial t} = 2A \cdot \frac{\partial A}{\partial t}$. Hence the expression is equal to

$$\begin{aligned} \frac{c}{4\pi} \left(H \cdot \operatorname{curl} E - E \cdot \operatorname{curl} H + \frac{\epsilon}{c} E \cdot \frac{\partial E}{\partial t} + \frac{\mu}{c} H \cdot \frac{\partial H}{\partial t} \right) = \\ \frac{c}{4\pi} \left[H \cdot \left(\operatorname{curl} E + \frac{1}{c} \frac{\partial B}{\partial t} \right) - E \cdot \left(\operatorname{curl} H - \frac{1}{c} \frac{\partial D}{\partial t} \right) \right]. \end{aligned}$$

But by Maxwell's equations $\operatorname{curl} E + \frac{1}{c} \frac{\partial B}{\partial t} = 0$, $\operatorname{curl} H - \frac{1}{c} \frac{\partial D}{\partial t} = \frac{4\pi u}{c}$, so that the result is $-E \cdot u$. Hence Poynting's theorem is

$$\operatorname{div} \left[\frac{c}{4\pi} (E \times H) \right] + \frac{\partial}{\partial t} \left[\frac{1}{8\pi} (\epsilon E^2 + \mu H^2) \right] = -E \cdot u. \quad (6)$$

From the analysis of the last section, we see that $-E \cdot u$ must represent the total rate of production of electromagnetic energy by e.m.f.s minus the rate of dissipation into heat. The latter is simple: in regions where Ohm's law holds, $u = \sigma E$, so that here we have the contribution $-\sigma E^2$ to the right side. The quantity σE^2 represents the ordinary dissipation of energy into heat. We must examine the other sort of term, the external e.m.f., rather more carefully.

159. The Nature of an E.M.F.—In a conductor carrying a current, there will be a current u set up, equal to the total force

per unit charge, times σ . The force is ordinarily simply the electrical force E . But sometimes there are other sorts of force acting. For example, in a battery, the various concentrations of electrolytes produce a definite pressure on the ions, forcing them mechanically in one direction, and this force would not ordinarily be considered as being electrical in nature. Inside a battery, the electric field is actually opposite to the flow of current, pointing from positive pole to negative, while the current flows from negative to positive. But the additional force acting on the charges counteracts the electric field, and does enough more so that it can push the current through the internal resistance of the battery. This latter part is already taken care of in computing the work done by the resistance. The former part, just equal and opposite to the E in the battery, is the force responsible for the applied e.m.f. of the battery. Thus it is $-E$ per unit charge. And the rate of working of the force on unit charge is the force times the velocity of the charge. We actually wish the rate of working per unit volume, so that we must multiply by the charge per unit volume. This is ρ , and its product with the velocity is just the current density u , so that we have $-E \cdot u$ as the rate of working of the e.m.f. on the electrical system. This is just the contribution to the right side of Poynting's theorem which we should get inside the batteries.

160. Examples of Poynting's Vector.—The conception of the energy of the electromagnetic field as residing in the medium is a very fundamental one, which has had great influence in the development of the theory. Thus Maxwell thought of the medium as resembling an elastic solid, the electrical energy representing the potential energy of strain of the medium, the magnetic energy the kinetic energy of motion. Such a definite view is no longer held. Nevertheless, the energy is always believed to travel through space. Thus, in a light wave, there is a certain energy per unit volume, proportional to the square of the amplitude (E or H). This energy travels along, and Poynting's vector is the vector which measures the rate of flow, or the intensity of the wave. We shall show that the vector actually points along the ray of light, the direction of flow. If, for example, we have a source of light, and we wish to find at what rate it is emitting energy, we surround it by a closed surface, and integrate the normal component of Poynting's vector over the surface. The whole conception of energy being transported in the medium is

evidently quite fundamental to the electromagnetic theory of light.

When we come back to charges and currents, however, it is a little harder to see the significance of the energy in the medium. For example, in a circuit consisting of a battery, and a wire connecting the plates, Poynting's vector indicates that the energy flows out of the battery, through the space surrounding the wire, and finally flows into the wire at the point where it will be transformed into heat. This seems to have small physical significance. In a moving electron, the situation is somewhat more reasonable. Suppose that the electron at rest is to be represented by a sphere of radius R , on the surface of which the charge is distributed. Then the field will be e/r^2 at any point outside the sphere. The total electrical energy is the volume integral of $\frac{1}{8\pi} \frac{e^2}{r^4}$ over all space outside the sphere, or

$$\frac{1}{8\pi} e^2 \int_R^\infty \frac{4\pi r^2}{r^4} dr = \frac{e^2}{2R}.$$

In the theory of the electron, it is this quantity which is interpreted as being the actual constitutive energy of the electron; although a correction must be made of an additional energy required to keep the sphere in equilibrium. Neglecting this correction, we can compute the mass of the electron. For a relation of Einstein says that a given energy has a mass, given by the relation, energy = mc^2 . Hence $mc^2 = e^2/2R$. Solving for the radius, we have $R = e^2/2mc^2$, a familiar formula for the radius of the electron. The correct formula, inserting the correction we omitted, differs only by a small factor. Inserting the correct values of $e = 4.774 \times 10^{-10}$ e.s.u., $m = 9.00 \times 10^{-28}$ gm., $c = 3 \times 10^{10}$ cm. per second, we have $R = 1.41 \times 10^{-13}$ cm. Now if this electron moves, it will have a magnetic field, as a current would, and hence will have a certain magnetic energy. Since the magnetic field is proportional to the velocity (or the current), the magnetic energy is proportional to the square of the velocity. This can be shown to be the kinetic energy. Further, there will be a Poynting vector, pointing in general in the direction of travel of the electron, and representing the flow of energy associated with the electron. All these relations prove on closer examination to be more complicated than they seem at first sight; but they lead to a consistent theory of the nature of the electron.

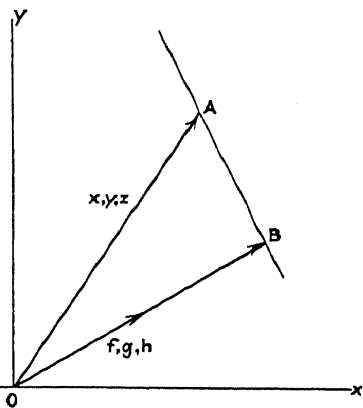
It should be stated, however, that this theory does not fit in with the quantum theory, and that its correct form on the basis of that theory is not known at present.

161. Energy in a Plane Wave.—Let us compute the flow of energy in a plane wave of light. It was shown in the last chapter and its problems that the potentials and fields satisfy a wave equation of the form

$$\nabla^2 E - \frac{\epsilon\mu}{c^2} \frac{\partial^2 E}{\partial t^2} = 0, \quad (7)$$

corresponding to propagation with the velocity $v = c/n$, where $n = \sqrt{\epsilon\mu}$. Here n , the ratio of the velocity of light in empty space to the velocity in the medium we are interested in, is the index of refraction. It is easy to set up a plane wave solution of the wave equation. Thus a wave of frequency ν , propagated in a direction whose direction cosines are f, g , and h , is represented by

$$E = E_0 e^{2\pi i \nu \left[t - \frac{n(fx + gy + hz)}{c} \right]}. \quad (8)$$



E_0 is a constant vector, measuring the amplitude of the wave. The exponent is constant, representing constant phase, or a wave front, when $fx + gy + hz = (c/n)t = vt$. Now $fx + gy + hz$ is the projection of the radius vector x, y, z on the direction f, g, h , so that, as we see from Fig. 42, all points for which $fx + gy + hz$ is constant lie on a plane whose normal is f, g, h , and whose distance from the origin is given by the constant. If this constant is vt , the plane travels out with a velocity v , as a wave front should. To have a wave of arbitrary phase, E_0 would have to be a complex vector. We can immediately show by substitution that the wave as we have written it is a solution of the wave equation. For instance, $\partial E / \partial x = -(2\pi i \nu n f / c) E$, and carrying out the various differentiations and substitutions, and making use of the relation $f^2 + g^2 + h^2 = 1$, the result follows at once.

Having the form of the solutions for E and H , we may apply Maxwell's equations. We note that the wave equations separate

E and H completely, but Maxwell's equations prescribe relations between them, so that actually Maxwell's equations are more restrictive than the wave equation. First, we cannot hope to satisfy the relations unless E and H both have the same exponential factor, corresponding to the same frequency and wave normal. Assuming this to be true, we can apply the equations in succession. Let us first take $\text{div } D = 0$. This leads at once to $-\frac{2\pi i \nu n \epsilon}{c}(fE_x + gE_y + hE_z) = 0$, showing that the scalar

product of unit vector along the wave normal, which we may call κ , and E , is zero. In other words, E and D have no component along the wave normal, or are in the plane of the wave front. Similarly $\text{div } B = 0$ shows that B and H are in the plane of the wave front. Next take the curl equations, beginning with $\text{curl } H = \frac{\epsilon}{c} \frac{\partial E}{\partial t}$. This gives for its x component

$$-\frac{2\pi i \nu n}{c}(gH_z - hH_y) = \frac{\epsilon}{c}(2\pi i \nu)E_x,$$

which is the x component of

$$E = -\frac{n}{\epsilon}(\kappa \times H), = -\sqrt{\frac{\mu}{\epsilon}}(\kappa \times H),$$

showing that E is at right angles both to H and the wave normal, these three then forming a set of three orthogonal directions. Further, since κ and H are at right angles, the magnitude of E equals $\sqrt{\mu/\epsilon}$ times the magnitude of H . The fourth equation can be easily shown to lead to the same condition.

Now we find the energy density. It is evidently

$$\frac{1}{8\pi}(\epsilon E^2 + \mu H^2) = \frac{\epsilon E^2}{4\pi},$$

as we see from the relations between E and H . Setting $E = E_0 \cos 2\pi \nu \left[t - \frac{n}{c}(fx + gy + hz) \right]$, and squaring, we have a quantity oscillating with time, but its time average, which alone has physical significance, is $E_0^2/2$. Hence the mean energy density is $\epsilon E_0^2/8\pi$. Next, Poynting's vector, being at right angles to E and H , is along κ , as it should be. Its magnitude is $(c/4\pi)E \times \sqrt{\epsilon/\mu}E$, so that its mean is $(c/8\pi)\sqrt{\epsilon/\mu}E_0^2$, or $c/\sqrt{\epsilon\mu}$ times the energy density. But this is the result we

should expect. This energy would be contained in a volume 1 sq. cm. in cross section, and of length $v = c/\sqrt{\epsilon\mu}$ cm. But if the light moves with a velocity v along the long axis of the volume, this energy will cross the 1 sq. cm. in one second, so that it should represent the flow vector, or Poynting's vector.

162. Plane Waves in Metals.—Let us consider the propagation of a plane wave in a metallic conductor, where for simplicity we shall take $\mu = 1$, $\rho = 0$, but $u = \sigma E$. Rather than satisfying the wave equation first and then substituting in Maxwell's equations, as we did in the preceding case, we shall vary the procedure by assuming a wave with undetermined velocity, and satisfying all four of Maxwell's equations (in the preceding case only three of Maxwell's equations, and the wave equation, were actually used, Maxwell's fourth equation being automatically satisfied). Let us then assume that E and H are given by expressions of the form

$$E_0 e^{2\pi i \nu \left[t - \frac{a}{c}(fx + gy + hz) \right]}, \quad (9)$$

where a is to be determined. The divergence equations show as before that E and H are both in the plane of the wave front. The equation for curl E leads to $a(\kappa \times E) = H$, showing as before that E and H are at right angles to each other, and that the magnitude of H is a times the magnitude of E . The equation $\text{curl } H = \frac{\epsilon}{c} \frac{\partial E}{\partial t} + \frac{4\pi\sigma}{c} E$ gives a new condition,

$$-\frac{2\pi i \nu a}{c} (\kappa \times H) = \left(\frac{\epsilon}{c} 2\pi i \nu + \frac{4\pi\sigma}{c} \right) E.$$

This condition likewise shows that E and H are at right angles to each other, but now gives the magnitude of H equal to $\frac{1}{a} \left(\epsilon - \frac{2i\sigma}{\nu} \right)$ times the magnitude of E . These conditions are only consistent if

$$a = \frac{1}{a} \left(\epsilon - \frac{2i\sigma}{\nu} \right), \quad a^2 = \epsilon - \frac{2i\sigma}{\nu}. \quad (10)$$

We see, in other words, that a , the quantity corresponding to the index of refraction, is complex. Let us write $a = n - ik$, where n and k are real, so that, as we can easily see,

$$n^2 - k^2 = \epsilon, \quad nk = \sigma/\nu,$$

and

$$\begin{aligned} n &= [\tfrac{1}{2}(\sqrt{\epsilon^2 + 4\sigma^2/\nu^2} + \epsilon)]^{\frac{1}{2}}, \\ k &= [\tfrac{1}{2}(\sqrt{\epsilon^2 + 4\sigma^2/\nu^2} - \epsilon)]^{\frac{1}{2}}. \end{aligned} \quad (11)$$

To understand the meaning of n and k , we substitute in the original expression for the plane wave, Eq. (9). This can be written

$$E_0 e^{-\frac{2\pi\nu k}{c}(fx+gy+hz)} e^{2\pi i\nu \left[t - \frac{n}{c}(fx+gy+hz) \right]}.$$

The second factor is just like an ordinary plane wave, with index of refraction n , though since n depends on frequency, we find the Maxwell theory predicting dispersion of electromagnetic waves in metals. But the first factor, a pure exponential term decreasing as $fx + gy + hz$ increases, means that there is a decrease of amplitude and energy as the wave travels along, or an absorption, as we can easily see from an application of Poynting's theorem, computing the Joule heating within the metal. For this reason k is called the absorption coefficient.

We have found that the magnitude of H is a , or $n - ik$, times the magnitude of E . If we write the complex number $n - ik$ in the exponential form, we have

$$n - ik = \sqrt{n^2 + k^2} e^{-2\pi i\nu\delta},$$

where $\delta = \frac{1}{2\pi\nu} \tan^{-1} \frac{k}{n}$, and

$$|H| = E_0 \sqrt{n^2 + k^2} e^{-\frac{2\pi\nu k}{c}(fx+gy+hz)} e^{2\pi i\nu \left[t - \frac{n}{c}(fx+gy+hz) - \delta \right]},$$

so that there is a phase difference between E and H in a conductor, whereas in an insulator they are in phase. The details of the calculation of electric and magnetic energies are left to a problem.

Problems

1. If the generation of heat per cubic centimeter in a conductor carrying a current is σE^2 , prove that for a cylindrical conductor of resistance R , carrying a current i , the rate of generation is $i^2 R$.

2. Given a cylindrical wire carrying a current. Find the values of E and H on the surface of the wire, computing Poynting's vector, and show that it represents a flow of energy into the wire. Show that the amount flowing into a given length of wire is just enough to supply the energy which appears as heat in the length. Note that the surface of a wire carrying current is not an equipotential so that there can be a component of electric field parallel to it.

3. Prove $\text{div} (A \times B) = B \cdot \text{curl} A - A \cdot \text{curl} B$.

4. The maximum electric field in a light wave is 0.1 volt per centimeter. Find how much energy is transported by the beam across 1 sq. cm. per second.

5. Given a 40-watt lamp, and suppose that all its energy is dissipated in radiation of one wave length or another. Take a sphere of radius 1 m. surrounding it, and suppose the radiation is of equal intensity in all directions. Find the maximum electric field in the radiation at this distance, in volts per centimeter, and the maximum magnetic field in gauss. Find the energy per cubic centimeter at this distance, in ergs per cubic centimeter.

6. Apply Poynting's theorem to the case of a plane wave traveling in a conductor and show that the rate of dissipation of electromagnetic energy just equals the Joule heating.

7. Calculate the electric and magnetic energies in a plane wave traveling in a metal and show by direct comparison that they are different from each other. What happens in the limiting cases $\sigma \rightarrow 0$ and $\sigma \rightarrow \infty$, *i.e.*, insulators and perfect conductors?

8. Investigate the behavior of n and k for a metal as functions of frequency, drawing curves. Take $\epsilon = 1$, and take the conductivity of copper. Note that the conductivity in electrostatic units has the dimensions of a frequency, and find in what part of the spectrum this frequency lies. Show that the value of ϵ is only significant when the frequency becomes greater than σ .

9. The significance of σ as a frequency is found from the relaxation time, the time taken for a volume charge set up within a metal to die down to $1/e$ th of its original value. Derive this in the following manner. Set up the equation of continuity for the current density u and charge density ρ . In this, write u in terms of E by Ohm's law, and write the result in terms of ρ by the relation $\epsilon \operatorname{div} E = 4\pi\rho$. Solve the resulting differential equation for ρ , showing that the solution is $\rho = \rho_0 e^{-t/\tau}$, where τ , the relaxation time, is $\epsilon/4\pi\sigma$, so that σ is, as far as its order of magnitude is concerned, the frequency connected with the relaxation time.

CHAPTER XXIII

REFLECTION AND REFRACTION OF ELECTROMAGNETIC WAVES

According to the electromagnetic theory of light, light consists of electromagnetic waves, propagated according to Maxwell's equations. We have already seen how we are led to the wave equation for E and H , or for the potentials, and we have investigated the plane wave solutions of these equations, showing that E and H are at right angles to each other and to the direction of propagation, the latter being the same as the direction of Poynting's vector, giving the energy flow. We shall now investigate the electromagnetic theory of some simple optical phenomena, beginning with reflection and refraction.

163. Boundary Conditions at a Surface of Discontinuity.—We have seen in the last chapter the conditions that hold for a wave in a refracting medium, whose index of refraction is constant. In the problem of reflection and refraction at a boundary between two media, however, the index changes suddenly from one medium to the other, and we must investigate what happens there. Let us assume that the boundary is a plane normal to the z axis. Then we shall apply Maxwell's equations, in the integrated form, to small regions containing the boundary. Thus take a thin flat volume, its faces parallel to the boundary and containing it. Let the area of the face be A . Apply to the above the divergence theorem, $\text{div } D = 4\pi\rho$, or $\iiint \text{div } D \, dv = \iint D_n dS = 4\pi q$, where q is the total charge within the volume. The surface integral comes almost wholly from the flat faces; it is $A(D_{n2} - D_{n1})$, if D_2 is the value of D in the upper medium, D_1 in the lower. If now the surface is uncharged, q gets smaller and smaller as the volume becomes thinner, so that in the limit $A(D_{n2} - D_{n1}) = 0$, or $D_{n2} = D_{n1}$. That is, the normal component of D is continuous at an uncharged surface.

Next let us apply the curl equations, to contours of the following sort: infinitesimal contours of long thin shape, in which one long side is in one medium, the other in the other, parallel to

the surface, and the parts of the contour which cross over from one medium to the other are of negligible length compared with the long sides. Consider $\text{curl } H = \frac{1}{c} \frac{\partial D}{\partial t} + \frac{4\pi u}{c}$, or integrated,

$$\int H_s ds = \int \int \left(\frac{1}{c} \frac{\partial D_n}{\partial t} + \frac{4\pi u_n}{c} \right) dS.$$
 If there is no surface current, D and u are finite vectors, so that as the contour gets narrower and narrower, and the area smaller and smaller, the right side of this equation will vanish. The left side approaches $(H_{s2} - H_{s1})L$, where L is the length of the contour, H_{s1} and H_{s2} are the tangential components of H in the media 1 and 2, respectively. Thus finally we have $H_{s2} = H_{s1}$, or the tangential component of H is continuous. Similarly we show that the tangential component of E is continuous.

Now we can see how to solve a problem involving two media separated by a plane surface, as air and glass. In one medium, we assume a plane wave approaching the boundary. But it must stop at the boundary, for the same plane wave, with the same wave length, would not be a solution of the problem for the second medium. There must be some wave in the second medium, however, for otherwise the boundary conditions could not be satisfied. Thus we are led to the existence of the refracted ray. As a matter of fact, we find that we cannot satisfy the boundary conditions without an incident, refracted, and also a reflected ray. By using all these, with proper relations between direction, amplitude, etc., we can actually satisfy the boundary conditions at the surface of separation of two media.

164. The Laws of Reflection and Refraction.—Assume a plane wave in the first medium, striking the surface of separation.

This wave will have the form $e^{2\pi i \nu \left(t - \frac{lx + my + nz}{v} \right)}$. Let the surface of separation be given by $z = 0$, the xy plane. Further let the axis be so chosen that the wave normal is in the xz plane, as in Fig. 43, so that $m = 0$. Then at points of the surface of separation the disturbance is given by $e^{2\pi i \nu \left(t - \frac{lx}{v} \right)}$. It is this disturbance which, taken together with the corresponding expressions from the reflected and refracted waves, must satisfy certain boundary conditions.

Next we consider a possible refracted wave. It will be in general of the form $e^{2\pi i \nu' \left(t - \frac{l'x + m'y + n'z}{v'} \right)}$, so that in the surface

of separation it will reduce to the value of this with $z = 0$. The boundary conditions must be satisfied for all values of x, y , and t , and yet we have only one constant at our disposal, an amplitude, in addition to the frequency and direction. It is obvious that the only possibility of satisfying the conditions will come if we make $\nu' = \nu$, $l'/v' = l/v$, $m' = 0$. For then we shall have just the same function of x, y , and t for both incident and refracted waves, at all points of the boundary. First,

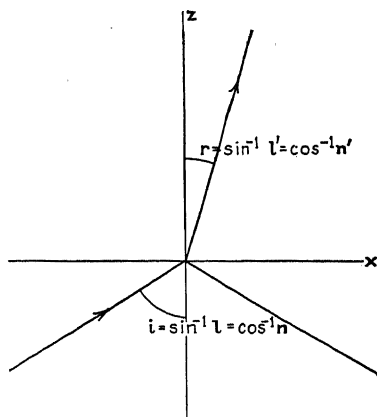


FIG. 43.—Law of refraction.

then, the refracted wave must have the same frequency as the incident one. Next, if the incident wave normal is in the xz plane, this must also be true of the refracted wave. Finally, there is a relation between the angle of incidence and the angle of refraction. We have $l = \cosine$ of the angle between the wave normal and the x axis = \sin of the angle between the wave normal and the normal to the surface = $\sin i$, where i is the angle of incidence. Similarly, $l' = \sin r$, where r is the angle of refraction. Thus we have

$\frac{l}{l'} = \frac{\sin i}{\sin r} = \frac{v}{v'} = \text{index of refraction of the second medium with respect to the first.}$ In other words, we have the ordinary law of refraction, as a necessary consequence of the boundary conditions.

Similarly, for the reflected wave, moving in the first medium, we see that m must be equal to zero, and l equal to the value for the incident wave, showing that the angle of reflection equals the angle of incidence. Now the reflected wave must be different from the incident wave, and to do this we must have the n for the reflected wave the negative of the value for the incident one, showing that the reflected wave travels away from the surface rather than towards it.

165. Reflection Coefficient at Normal Incidence.—After proving the laws of reflection and refraction, we still have much more to do to apply the boundary conditions. For we must compute the values of the various vectors at the surface, and actually satisfy the conditions. Let us take first the simple

case of normal incidence, where $l = 0$, and all waves travel along the z axis. Let us suppose that in the incident beam we have E along the x axis, H along y . For simplicity we assume the first medium to have the index of refraction unity, the second the index $n = \sqrt{\epsilon}$. Then in the refracted wave we assume that E is along the x axis, H along y , and that the value of E is E' , so that $H' = nE'$. In the reflected wave, assume that E has a changed phase, H not, so that E is along $-x$, H along y , and each numerically equal to E'' . The change of phase of one vector and not the other is necessary to reverse the direction of the Poynting's vector.

Now we may apply the boundary conditions. All normal components are zero, so that these conditions are automatically satisfied. For the tangential component of E , we have $E - E'' = E'$; for the tangential component of H , $H + H'' = H'$. The latter is then $E + E'' = nE'$. Combining the two, we have

at once $E' = \frac{2E}{n+1}$ (by adding), and $E'' = \frac{E'(n-1)}{2}$ (by sub-

tracting), leading to $\frac{E''}{E} = \frac{n-1}{n+1}$. This gives us directly the

reflection coefficient at normal incidence. The ratio of reflected to incident intensity is proportional to the ratio of the squares of

the amplitudes, or $\frac{(n-1)^2}{(n+1)^2}$. This shows that the reflected

intensity is never so great as the incident, but that the ratio approaches closer and closer to unity as n becomes larger. It is interesting to compute the reflection coefficient for familiar substances. For instance, for glass, n is about 1.5, so that the coefficient is $(0.5/2.5)^2 = 1/25$, showing that only a few per cent of the intensity is reflected from a glass plate at normal incidence.

We can check the energy relations: the amount of energy brought to the surface per unit time in the incident wave should equal the amount carried away in the refracted and reflected

waves. The first is $\frac{c}{4\pi}(E \times H)$, whose magnitude is $\frac{c}{4\pi}E^2$. The

reflected energy is $\frac{c}{4\pi} \frac{(n-1)^2}{(n+1)^2} E^2$. The refracted intensity is

$\frac{c}{4\pi}(E' \times H') = \frac{c}{4\pi} n E'^2 = \frac{c}{4\pi} \frac{(4n)}{(n+1)^2} E^2$. The sum of the

refracted and reflected intensities is

$$\frac{c}{4\pi} \left[\frac{(n-1)^2 + 4n}{(n+1)^2} \right] E^2 = \frac{c}{4\pi} E^2,$$

equal to the incident intensity.

166. Fresnel's Equations.—Now we pass to Fresnel's equations, the extension of the last section to an arbitrary angle of incidence. Here, for the first time, we meet the question of polarization. The vector E is at right angles to the direction of propagation, but that does not fix the direction uniquely, and it is said that the wave is polarized in a particular direction if its electric vector points in that direction. Let us then consider the two extreme

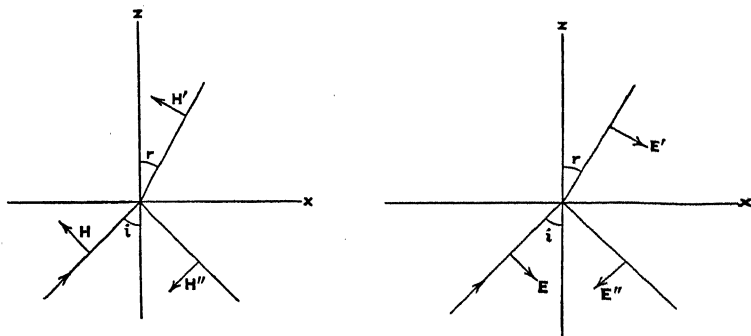


FIG. 44.—Vectors in reflection and refraction.

Case 1. y axis points down into the paper. E and E' point down, E'' points up.

Case 2. H, H', H'' all point down.

cases. We take the wave normal of the incident wave to be in the xz plane, as before. Then we consider the case where the electric vector is along the y axis, and the case where it is in the xz plane, as in Fig. 44.

CASE 1. Electric vector along the y axis. All vectors depend on space in the following way, rewriting l, m, n in terms of the angles of incidence and refraction: for the incident wave,

$$e^{2\pi i v \left(t - \frac{x \sin i + z \cos i}{v} \right)}, \quad (1)$$

for the refracted wave,

$$e^{2\pi i v \left(t - \frac{x \sin r + z \cos r}{v'} \right)}, \quad (2)$$

for the reflected wave,

$$e^{2\pi i v \left(t - \frac{x \sin i - z \cos i}{v} \right)}. \quad (3)$$

We take E and E' to be along the y axis. Then H is in the xz plane, at right angles to the wave normal. That is, for the incident wave, $H_x = -E \cos i$, $H_z = E \sin i$. Similarly, in the refracted wave, $H_x' = -nE' \cos r$, $H_z' = nE' \sin r$, and for the reflected ray $H_x'' = -E'' \cos i$, $H_z'' = -E'' \sin i$. Hence, we have the following relations:

Normal component of D : nothing, since D is tangential.

Normal component of B : $E \sin i - E'' \sin i = nE' \sin r$.

Tangential component of E : $E - E'' = E'$.

Tangential component of H : $-E \cos i - E'' \cos i = -nE' \cos r$.

Remembering that $\frac{\sin i}{\sin r} = n$, the first two equations reduce to the same equation, $E - E'' = E'$. The last is $E + E'' = \frac{n \cos r E'}{\cos i} = E' \frac{\tan i}{\tan r}$. From this at once, multiplying the first by $\frac{\tan i}{\tan r}$, and subtracting, we have

$$\begin{aligned} E \left(\frac{\tan i}{\tan r} - 1 \right) &= E'' \left(\frac{\tan i}{\tan r} + 1 \right), \\ \frac{E''}{E} &= \frac{\tan i - \tan r}{\tan i + \tan r} = \frac{\sin i \cos r - \cos i \sin r}{\sin i \cos r + \cos i \sin r}, \\ \frac{E''}{E} &= \frac{\sin(i - r)}{\sin(i + r)}. \end{aligned} \quad (4)$$

This gives the amplitude of the reflected wave, and is one of Fresnel's equations. We note that as i and r become zero, the law of reflection becomes $i/r = n$, $i = nr$. Thus in the limit of normal incidence, the ratio approaches $(nr - r)/(nr + r) = (n - 1)/(n + 1)$, as we found above. We also note, in the other extreme of tangential or grazing incidence, that $i = 90$ deg., so that the ratio is $\frac{\sin(90 \text{ deg.} - r)}{\sin(90 \text{ deg.} + r)} = 1$. That is, the reflection coefficient equals unity for grazing incidence. The formula gives a gradual increase of amplitude as the angle of incidence increases.

CASE 2. Electric vector in the xz plane. Let H be along the y axis in all the waves: $H_y = E$, $H_y' = nE'$; $H_y'' = E''$. Then we take $E_x = E \cos i$, $E_z = -E \sin i$, $E_x' = E' \cos r$, $E_z' = -E' \sin r$, $E_x'' = -E'' \cos i$, $E_z'' = -E'' \sin i$. Then we have:

Normal component of D : $-E \sin i - E'' \sin i = -n^2 E' \sin r$.

Normal component of B : nothing.

Tangential component of E : $E \cos i - E'' \cos i = E' \cos r$.

Tangential component of H : $E + E'' = nE'$.

Using the law of refraction, the first and last are the same,

$E + E'' = nE'$. The other is $E - E'' = E' \frac{\cos r}{\cos i}$. Multiplying

the first by $\frac{\cos r}{\cos i}$, the second by $n = \frac{\sin i}{\sin r}$, and subtracting, we have

$$E \left(\frac{\cos r}{\cos i} - \frac{\sin i}{\sin r} \right) = E'' \left(-\frac{\cos r}{\cos i} - \frac{\sin i}{\sin r} \right),$$

or

$$\frac{E''}{E} = -\frac{\cos r \sin r - \cos i \sin i}{\cos r \sin r + \cos i \sin i}.$$

Now we see at once that

$$\sin(i \pm r) \cos(i \mp r) =$$

$$(\sin i \cos r \pm \cos i \sin r)(\cos i \cos r \pm \sin i \sin r) =$$

$$\sin i \cos i (\cos^2 r + \sin^2 r) \pm \sin r \cos r (\sin^2 i + \cos^2 i) =$$

$$\sin i \cos i \pm \sin r \cos r.$$

Hence we have

$$\frac{E''}{E} = \frac{\sin(i - r) \cos(i + r)}{\sin(i + r) \cos(i - r)} = \frac{\tan(i - r)}{\tan(i + r)}. \quad (5)$$

This is the other of Fresnel's equations.

167. The Polarizing Angle.—In Case 2 of Sec. 166, where the electric vector is in the plane of incidence, or the xz plane, we notice an interesting fact. If $i + r = 90$ deg., a perfectly possible situation, we have $\tan(i + r) = \infty$, so that $E''/E = 0$. That is, the amount of reflected light, at this angle, is zero. There is no such situation for the other sort of polarization. Suppose, then, that we take an unpolarized beam, such as would be emitted by any ordinary source, and reflect it from a mirror at this angle, called the polarizing angle. The reflected light will consist entirely of the light polarized with the electric vector at right angles to the plane of incidence. It was by this phenomenon that polarized light was first discovered. Light was reflected from one mirror at this angle. Then its polarization was found by reflecting from a second mirror at the same angle. As the second mirror was rotated about the beam as an axis, so that the polarization changed from being at right angles to

the plane of incidence to being in the plane, the doubly reflected beam changed from a maximum intensity to zero.

The polarizing angle r' is fixed by $i' + r' = 90$ deg., and this occurs when $\cos i' = \sin r'$. Using the law of refraction, we find $\tan i' = n$, thus fixing the definite angle i' . For glass the angle of polarization is 56 deg.

168. Total Reflection.—For light passing from a dense medium with index of refraction n to a vacuum of index 1, the law of refraction is $n \sin i = \sin r$. For the angle of incidence given by $\sin i = 1/n$, we have $\sin r = 1$, $r = 90$ deg., and the refracted ray emerges at grazing incidence. For larger angles of incidence, $\sin r$ is greater than 1, and there is no real angle r . Physically we know that at these angles, greater than the critical angles, there is total reflection, with no transmitted beam. We can easily investigate the situation mathematically.

In the first place, let us consider the disturbance in the second medium, for we find there is a disturbance, even though no transmitted beam is observed. This is given by an exponential

$$e^{2\pi i\nu\left(t - \frac{x \sin r + z \cos r}{c}\right)},$$

where we remember that the second medium has index 1, velocity c . But $\cos r = \pm \sqrt{1 - \sin^2 r} = \pm \sqrt{-1} \sqrt{n^2 \sin^2 i - 1}$, a pure imaginary. Thus the exponential becomes

$$e^{2\pi i\nu\left(t - \frac{x \sin r}{c}\right)} e^{-2\pi\nu \frac{\sqrt{n^2 \sin^2 i - 1}}{c} z}, \quad (6)$$

where we have used the negative square root. The first term represents a wave propagated along the x axis, or parallel to the surface of the medium, with an apparent velocity $c/\sin r$, a value less than c . The second factor indicates that the amplitude of this wave is damped out as z increases, or as we go away from the surface, so that the wave fronts (surfaces of constant phase) are at right angles to the surfaces of constant amplitude. This disturbance ordinarily damps out in a very short distance. Thus if $n^2 \sin^2 i$ is decidedly greater than 1, the exponential becomes small when z is a few wave lengths ($\nu z/c$ a reasonably large number). Consequently the disturbance is not observed. It is easily shown that Poynting's vector for this wave has no component normal to the surface, so that it does not carry any energy away.

The reflected wave may be treated by Fresnel's equations. Thus, in Case 1 we have

$$\frac{E''}{E} = -\frac{\cos i \sin r - \sin i \cos r}{\cos i \sin r + \sin i \cos r} = -\frac{a - ib}{a + ib},$$

where $a = \cos i \sin r$, $b = -\sin i \sqrt{n^2 \sin^2 i - 1}$. This ratio can now be written as $-e^{-2i \tan^{-1} b/a}$, so that E'' and E are of the same magnitude, showing that all the light is reflected, but they differ in phase. We may write

$$\frac{E''}{E} = -e^{i\delta_1},$$

where

$$\tan \frac{\delta_1}{2} = \frac{\sqrt{\sin^2 i - 1/n^2}}{\cos i}.$$

Similarly for Case 2 we have

$$\frac{E''}{E} = \frac{\cos i \sin i - \cos r \sin r}{\cos i \sin i + \cos r \sin r} = \frac{c - id}{c + id},$$

where $c = \cos i \sin i$, $d = -\sin r \sqrt{n^2 \sin^2 i - 1}$. Again all the light is reflected, but with a change of phase δ_2 given by

$$\frac{E''}{E} = e^{i\delta_2},$$

where

$$\tan \frac{\delta_2}{2} = \frac{n^2 \sqrt{\sin^2 i - 1/n^2}}{\cos i}.$$

Thus, in the general case, where E has components both in the xz plane and along the y axis, there is a difference of phase between these components upon total reflection, and linearly polarized light in general will become elliptically polarized upon total reflection. To see this, we note that two vibrations at right angles, with the same frequency and phase, produce a resultant vector whose extremity moves in a line (plane polarization), but if the two components are in different phases the extremity of the vector traces out an ellipse. If the phases differ by 90 deg., and the amplitudes are equal, the polarization is circular.

It follows from our expressions for δ_1 and δ_2 that the difference between these phase angles, which we denote by δ , is given by the relation

$$\tan \frac{\delta}{2} = \frac{\cos i \sqrt{\sin^2 i - 1/n^2}}{\sin^2 i}.$$

Only in the case of grazing incidence, $i = \pi/2$, does $\delta = 0$, so that our above remarks hold valid except in this case. It is clear that by causing an elliptically polarized beam to be totally reflected at the correct angle, it can be transformed into a beam of linearly polarized light.

169. The Optical Behavior of Metals.—We shall now examine the law of reflection for light falling on metals, restricting the discussion to the case of normal incidence. In the last chapter we have already shown that in the case of metals we must introduce a “complex index of refraction,” $n' = n - ik$, where k is the extinction coefficient, and in so doing we retain the identical form of the relations which we have been using in this chapter. We have already found ($\mu = 1$)

$$n = \sqrt{\frac{1}{2}(\sqrt{\epsilon^2 + 4\sigma^2/\nu^2} + \epsilon)}$$

$$k = \sqrt{\frac{1}{2}(\sqrt{\epsilon^2 + 4\sigma^2/\nu^2} - \epsilon)}$$

where σ is the conductivity, ν the frequency, and ϵ the dielectric constant. ϵ is unknown for metals, but since σ for metals is so large (in e.s.u. $\sigma \cong 10^{18}$), we can neglect ϵ at least for light of sufficiently long period. Thus we find

$$n = k = \sqrt{\sigma/\nu} \quad (7)$$

relations first found by Drude. For the infra-red, $\sqrt{\sigma/\nu} \gg 1$.

We may still use Fresnel's equations as we have done for total reflection. For normal incidence these are simply

$$\frac{E''}{E} = \frac{n' - 1}{n' + 1},$$

and we must insert a complex value of n' for reflection from metals. Thus we have

$$E'' = E \frac{n - 1 - ik}{n + 1 - ik}$$

and taking the square, we find for the ratio of the reflected to the incident intensities,

$$R = \frac{n^2 + k^2 - 2n + 1}{n^2 + k^2 + 2n + 1} = \frac{(n - 1)^2 + k^2}{(n + 1)^2 + k^2}. \quad (8)$$

R is known as the reflective power of the metal. Since $n = k$, we may write

$$R = 1 - \frac{4n}{2n^2 + 2n + 1}$$

and since $n = \sqrt{\sigma/\nu} > 1$, this becomes

$$R = 1 - \frac{2}{n} = 1 - \frac{2}{k}$$

or

$$R = 1 - \frac{2}{\sqrt{\sigma/\nu}}. \quad (9)$$

This relation holds experimentally in the far infra-red, down to $\lambda \cong 5\mu$. The reflective power varies with the color of the incident light, and colors which are strongly absorbed are also strongly reflected.

Problems

1. Light is reflected from glass of index of refraction 1.5. Compute and plot curves for the reflected intensity as a function of angle, for both sorts of plane polarization.
2. Find the intensity of light in the refracted medium, for arbitrary angle of incidence and both types of polarization. Show that the amount of energy striking the surface is just equal to the amount carried away from it. Note that the amount striking the surface is computed, not from the whole of Poynting's vector, but from its normal component.
3. Show that the reflection coefficient from glass to air at normal incidence is the same as for air to glass, but that the phases of the reflected beams are opposite.
4. Light passes normally through a glass plate. Find the weakening in intensity on account of the reflection at the faces.
5. Ten plates of glass of index 1.5 are placed together and used as a polarizer. Light strikes the plates at the polarizing angle, and the transmitted light is used. Since all the reflected light is of one polarization, and the reflections at both surfaces of all plates are enough to remove practically all of the light of this polarization, the transmitted light will be practically polarized in the other direction. Find the intensity of both sorts of light in the transmitted beam, assuming initially unpolarized light, and hence show how much polarization is introduced. You may have to consider multiple internal reflection.
6. Derive the expressions for $\tan \delta_1/2$ and $\tan \delta_2/2$ in the paragraph on total reflection.
7. Derive the formulas for the phase difference δ of the two reflected components of E in the case of total reflection.
8. The conductivity of copper in e.s.u. is 5×10^{17} per second. Calculate the reflective power of copper for wave lengths of light $\lambda = 12\mu$ and $\lambda = 25.5\mu$. The observed values of $1 - R$ are 1.6 per cent and 1.17 per cent at these wave lengths.
9. Consider light linearly polarized so that the incident electric vector has equal components in the plane of the wave normal and the perpendicular

thereto. If this light falls on a metal, using Fresnel's equations find the ratio of the reflected components of E . If this ratio is written as $\rho e^{i\delta}$ show that

$$\frac{1 - \rho e^{i\delta}}{1 + \rho e^{i\delta}} = \frac{\sin i \tan i}{\sqrt{n'^2 - \sin^2 i}},$$

where i is the angle of incidence and n' the complex index of refraction of the metal.

CHAPTER XXIV

ELECTRON THEORY AND DISPERSION

Maxwell's theory and Maxwell's equations are based on the assumption of dielectrics with dielectric constant ϵ , magnetic substances with permeability μ , conductors with conductivity σ . These assumptions are unsatisfactory for two reasons. First, cases are known, and in fact are usual rather than exceptional, in which the three constants mentioned are not really constants. Thus the permeability of iron depends on the field strength. The dielectric constant of almost all substances depends on the frequency; as we have seen, the index of refraction n is given by the relation $n = \sqrt{\epsilon}$, and the well-known phenomenon of dispersion shows a dependence of refractivity on wave length or frequency. An extreme case is water, whose index of refraction in the visible is about 1.4, and whose dielectric constant is 80, a result of the fact that the dielectric constant is measured for static fields, and that n as a function of frequency goes from $\sqrt{80}$ at zero frequency, through a region in short radio or long infra-red waves in which the index greatly decreases, so that with the very high frequency of visible light it is reduced to 1.4. The second reason why Maxwell's assumptions are unsatisfactory is that, since matter is known to be composed of electric charges, electrons with negative charges and atomic nuclei with positive charges, it ought to be possible to explain these typically electrical properties of matter directly in terms of the electronic structure, without having to resort to empirical relations of the sort implied by a constant or variable dielectric constant. The attempt to derive the electrical properties of matter from the electron theory was first made by H. A. Lorentz, and he was successful not only in explaining the physical meaning of the dielectric constant, permeability, and conductivity, but in deriving their dependence on frequency, field strength, etc. Further developments of the theory, making use particularly of wave mechanics, have carried the subject much further than Lorentz was able to.

and in our later chapters on wave mechanics we return to these questions.

170. Polarization and Dielectric Constant.—The fundamental physical fact about a dielectric is that, when placed in an electric field, it acquires surface charges on its faces, proportional to the strength of the field. Thus in Fig. 45, a slab of dielectric is shown with positive and negative surface charges, as if the positive had actually been pulled along to the face by the action of the field, the negative pushed to the other face. These surface charges, of course, contribute to the field, just as do other charges, which we actually have control over.

The essence of the electron theory is that it treats these induced surface charges in the same way as any other charges, applying the ordinary Maxwell's equations to all charges in existence, and not considering dielectrics as being essentially different from free space, except in so far as they contain these polarizable electrons. Thus, if ρ and u are charge density and current density, respectively, of the so-called "real charge" which we can move about at will, and ρ_p and u_p the charge and current density of the charge arising from polarization, we assume Maxwell's equations for a nonmagnetic medium are

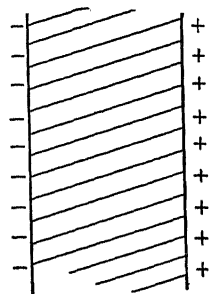


FIG. 45.—Polarization of dielectric.

$$\text{curl } H = \frac{1}{c} \frac{\partial E}{\partial t} + \frac{4\pi}{c}(u + u_p), \quad \text{curl } E = -\frac{1}{c} \frac{\partial H}{\partial t}$$

$$\text{div } E = 4\pi(\rho + \rho_p), \quad \text{div } H = 0. \quad (1)$$

In other words, we assume that the field E is the field of all charge, both "real" and polarization charge, and that the total current resulting from both sources produces the magnetic field.

The polarization charge must be produced, from the originally uncharged dielectric, by the motion of positive charges in the direction of E , and of negative charge in the opposite direction. Suppose that in equilibrium two equal charges of opposite sign lie so near together that they exert no appreciable external effect. By means of an external field these charges may be displaced relative to each other by a distance r . The charges then form a dipole of moment

$$p = er.$$

In producing such a dipole there is clearly a current

$$e \frac{dr}{dt} = ev = \frac{dp}{dt}.$$

If we add the dipole moments of all the polarization electrons in a unit volume we obtain the polarization vector, or the dipole moment per unit volume

$$P = \Sigma p, \quad (2)$$

and a current density due to these electrons equal to

$$u_p = \rho_p v_p = \frac{\partial P}{\partial t}. \quad (3)$$

In producing dielectric polarization, charges cross a surface in the body. In fact all the charges pass across the surface which originally were contained in a cylinder of base equal to the surface and length r . If r_n is the component of r normal to the surface, then we have as the charge passing through the end dS

$$\Sigma e r_n dS = P_n dS \quad (4)$$

which is the surface charge appearing on dS if this is an element of the outer surface of the body. If we consider a closed surface, the enclosed volume loses the charge

$$\iiint P_n dS = \iiint \text{div } P dv$$

according to Gauss's theorem. The density of polarization electrons remaining is given by $\rho_p = -\text{div } P$, since these have the opposite sign to those which have crossed the surface. We thus can write both polarization charge and current in terms of the polarization vector P .

We have seen that the field E is that resulting from all charge, including the polarization charge. The displacement D , however, is simply the field resulting from the real charge ρ , so that $\text{div } D = 4\pi\rho$. To get Maxwell's equations in terms of D , we take Eqs. (1), and make the substitutions

$$\begin{aligned} \rho_p &= -\text{div } P, \\ u_p &= \frac{\partial P}{\partial t}, \end{aligned} \quad (5)$$

which, as we note, obviously satisfy the continuity equation for polarization charge and current. Then we have at once, for the only two equations affected by the change,

$$\text{curl } H = \frac{1}{c} \frac{\partial}{\partial t} (E + 4\pi P) + \frac{4\pi u}{c} \quad (6)$$

$$\text{div} (E + 4\pi P) = 4\pi\rho.$$

If we set $D = E + 4\pi P$, these become the ordinary Maxwell equations.

171. The Relations of P , E , and D .—We have seen that E measures the field of all charge, D that due to the “real” charge, and that P is the polarization per unit volume. To understand P better, we may take a unit cube of dielectric, one pair of faces being perpendicular to the field. Since the polarization surface charge is P_n , one of these faces will have a charge on its unit area of $|P|$, the other of $-|P|$, so that the dipole moment of the cube, coming from these two charges at unit distance apart,

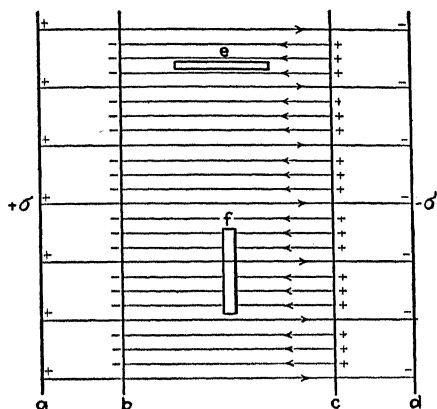


FIG. 46.—Condenser containing dielectric. Condenser plates a and d have surface charges $\pm\sigma$. Induced surface charges are shown on faces b and c of dielectric. The force on unit charge within cavity e is E , and within cavity f is D .

would be P . Similarly, if the volume had had length L parallel to the field, area A in the plane at right angles, the charges on the ends would be $\pm PA$, and the moment, remembering that these are a distance L apart, is PAL , or P times the volume, showing that the moment is proportional to the volume, so that it is really correct to regard P as the moment per unit volume.

The relations of the three quantities are perhaps best understood from a simple illustration in the theory of the condenser. In Fig. 46 we have a condenser consisting of two parallel plates a and d with surface charges $\pm\sigma$, respectively. Between them there is a slab of dielectric bc , with surface charges $\pm P$, on the faces c and b , respectively. The field E now is determined from

the whole charge; that is, using our relation regarding the relation of discontinuity of field to surface charge, the field within the dielectric is given by

$$E = 4\pi(\sigma - P).$$

The displacement D , however, is determined only from σ , so that

$$D = 4\pi\sigma = E + 4\pi P. \quad (7)$$

The capacity of the condenser is given by the charge, $D/4\pi$, divided by the potential difference, E times the distance L between the plates, or is $\frac{D}{E} \frac{1}{4\pi L}$. If we define the dielectric constant ϵ as the ratio D/E , this leads correctly to the relation that the capacity of the condenser is ϵ times the capacity of the same condenser with vacuum in place of the dielectric.

Let us now consider the meaning of the field within the dielectric. Actually, on account of the atomic and electronic structure, the field will change rapidly from point to point, so that it is not so easy as it might seem to define it. The usual method is to set up a long needle-shaped cavity e , pointing in the direction of the field. A point charge placed within the cavity would now be acted on by just the field of real and polarization charges, so that the field E is the force on unit charge in such a cavity. The necessity of choosing that particular shape of cavity is shown by considering the cavity f , which is supposed to be disk-shaped, with its flat face perpendicular to the field. This cavity will have surface charges $\pm P$ set up on its two faces, and it is evident that the lines of force starting from the polarization surface charges on plates b and c will terminate on these faces of the cavity, not crossing it at all, so that the field within it will come wholly from the real surface charges on a and d , or will be $E + 4\pi P = D$. Thus if we choose we may define E as the field in a cavity shaped like e , in which the effect of the charges on its faces is negligible because the faces are of negligibly small area and arbitrarily far from the point where we are finding the field, while we may define D as the field in a cavity shaped like f . These definitions were originally used for the corresponding magnetic case, by Kelvin. It is interesting to notice that the fields in cavities of other shapes are different, depending on the shape of the cavity. Thus in a later section we shall see that the field in a spherical cavity is $E + (4\pi P/3)$.

We notice finally that since $D = \epsilon E = E + 4\pi P$, we have $\epsilon = 1 + (4\pi P/E)$, a constant if the polarization is proportional to the field. To compute the dielectric constant, or refractive index, we have then to find the polarization per unit field, and we proceed to do this for gases, and later for solids.

172. Polarizability and Dielectric Constant of Gases.—In gases the molecules or atoms are relatively so far apart that we can neglect the interactions between them. Each molecule contains charges which can be displaced under the action of an external field, and these charges act as if they were held to positions of equilibrium by restoring forces proportional to the displacement. Thus in a static case an electron e is acted on by the forces eE of the external electric field, and $-cx$ the linear restoring force. The displacement is then $x = (e/c)E$, and the induced dipole moment $ex = (e^2/c)E$. The ratio e^2/c , giving the dipole moment set up by unit field, is called the polarizability, denoted by α . Thus the dipole moment per molecule is αE , and if there are N molecules per unit volume the polarization P is $N\alpha E$, so that $\epsilon = 1 + 4\pi N\alpha$.

A very simple model of an atom will give us the order of magnitude of the polarizability. The atom consists of a nucleus of charge Ze , where Z is an integer, e the magnitude of the charge on the electron, surrounded by a distribution of negative charge equal to $-Ze$. In the external field the negative charge will be displaced with respect to the nucleus. The restoring force may be computed as if the negative charge filled a sphere of radius R with uniform charge density. Then the positive charge Ze , at distance r from the center, would be acted on by a force as if the negative charge within a sphere of radius r were concentrated at the center, all other negative charge being neglected. This charge would be r^3/R^3 times the total charge, so that the force would be $\frac{(Ze)^2 r}{R^3}$. The polarizability is then found to be R^3 , proportional to the volume of the molecule.

173. Dispersion in Gases.—We now assume a sinusoidal external field of frequency ν , as in a light wave. The magnetic force on the electron on account of its motion can be neglected. In addition to the external electric force, and the elastic restoring force, we introduce a damping force proportional to velocity, to account for absorption. The equation of motion for the electron is then, for the x coordinate,

$$m\ddot{x} + m\gamma\dot{x} + \omega_0^2 mx = eE_x^0 e^{i\omega t} \quad (8)$$

where we have placed $\omega = 2\pi\nu$. Thus we have the problem of the damped linear oscillator in forced oscillation. We have solved this problem in Chap. IV, and can write for the steady-state solution

$$x = \frac{\frac{e}{m} E_x^0 e^{i\omega t}}{\omega_0^2 - \omega^2 + i\omega\gamma} = \frac{\frac{e}{m} E}{\omega_0^2 - \omega^2 + i\omega\gamma} \quad (9)$$

in complex form. This shows that the electron vibrates at the same frequency as the light wave but with an amplitude depending on the frequency and out of phase with the light wave. If we have N electrons per unit volume characterized by the constants ω_k and γ_k (electrons of the k th kind) we get for the dipole moment per unit volume:

$$P = \sum ex = E \sum_k \frac{N_k \frac{e^2}{m}}{\omega_k^2 - \omega^2 + i\omega\gamma_k}$$

whence we get for the displacement vector

$$D = E + 4\pi P = E \left(1 + 4\pi \sum_k \frac{N_k \frac{e^2}{m}}{\omega_k^2 - \omega^2 + i\omega\gamma_k} \right)$$

and if we introduce a "dynamic" refractive index $(n - ik)$ defined by $D = \epsilon E = (n - ik)^2 E$, we find

$$(n - ik)^2 = 1 + 4\pi \sum_k \frac{N_k \frac{e^2}{m}}{\omega_k^2 - \omega^2 + i\omega\gamma_k} \quad (10)$$

so that the index of refraction is a function of the frequency of the light, and different colors travel with different velocities. This is known as the dispersion of light. Furthermore, in general, the index of refraction is a complex quantity and, as we have seen in our discussion of electromagnetic waves in metals, this indicates absorption and is not surprising in view of our introduction of a damping force.

In the limit of slow frequencies (long wave lengths of light where $\omega \ll \omega_k$) we may neglect the last two terms in the denominator and find

$$n^2 = \epsilon = 1 + 4\pi \sum_k \frac{N_k \frac{e^2}{m}}{\omega_k^2}$$

as the static value of the dielectric constant of insulators, agreeing with the value found in the last section.

If the frequency of the light does not lie near any of the natural frequencies of the electrons, we may neglect the frictional force and find a real index of refraction given by

$$n = 1 + 2\pi \sum_k \frac{N_k \frac{e^2}{m}}{\omega_k^2 - \omega^2},$$

if we remember that the index of refraction for gases varies but slightly from unity. Thus there is no absorption and we have the case of normal dispersion. Let us consider the index of refraction as a function of frequency of the light in the visible region of the spectrum. If the natural frequencies of the electrons lie in the ultra-violet (and also for the case that they lie in the infra-red) the index of refraction increases with increasing frequency, the normal behavior.

In case the frequency of the light lies near a natural frequency, we obtain the phenomenon of anomalous dispersion. In this case the frictional term becomes important and we find an absorption band in the neighborhood of ω_0 . The whole discussion is similar to the case of a resonant electric circuit. For simplicity let us assume only one resonant frequency. Remembering that for gases n is very nearly unity, we have:

$$n - ik = 1 + 2\pi \frac{e^2}{m} \frac{N}{\omega_0^2 - \omega^2 + i\omega g}$$

and if we separate into real and imaginary parts, we obtain,

$$n = 1 + 2\pi \frac{e^2 N}{m} \frac{\omega_0^2 - \omega^2}{(\omega_0^2 - \omega^2)^2 + \omega^2 g^2} \quad (11)$$

and

$$k = 2\pi \frac{e^2 N}{m} \frac{\omega g}{(\omega_0^2 - \omega^2)^2 + \omega^2 g^2}. \quad (12)$$

n is known as the principal index of refraction and k the absorption coefficient. If we plot $n - 1$ and k against the light frequency, we get curves of the form shown in Fig. 47. Such

curves have already been considered in Prob. 10, Chap. IV. In the neighborhood of the absorption region we see that the index of refraction decreases with increasing frequency and this is the anomalous behavior giving rise to the term anomalous dispersion.

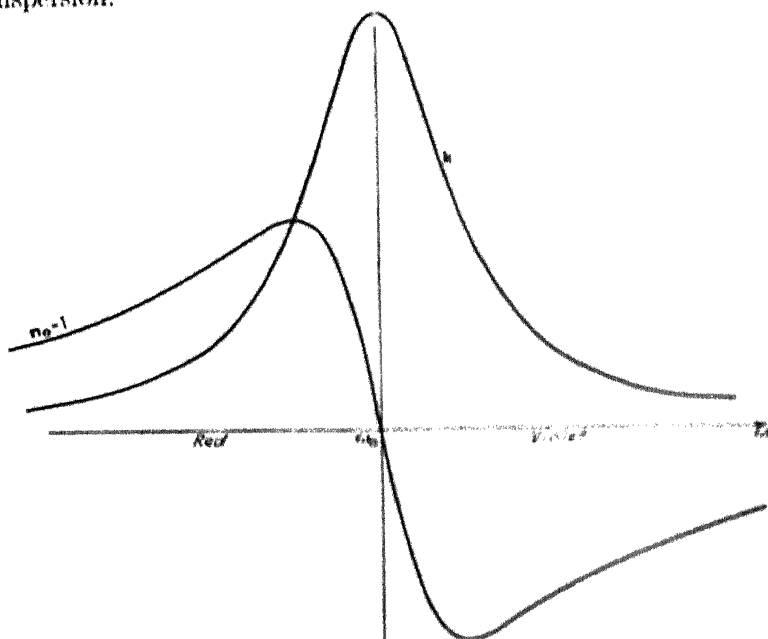


FIG. 47.—Anomalous dispersion, showing index of refraction and absorption coefficient as function of frequency.

174. Dispersion of Solids and Liquids.—In the case of solids and liquids we may no longer make the approximation that the force acting on an electron is simply the electric vector of the light wave in free space, but must take into account the added force on the electron due to the polarization of the body. We can calculate this force as follows: we imagine a small sphere of radius R (with its center at the position of the electron in question) cut out of the medium. If we do this, we have induced charges on the surface of this spherical volume from which we calculate the force at the center of the sphere. We have for the surface density of induced charge on a spherical ring at an angle θ , $\sigma = P_n = P \cos \theta$, as in Fig. 48. The area of the ring is $2\pi R \sin \theta \cdot R d\theta = 2\pi R^2 \sin \theta d\theta$, so that the charge on this ring is

$$2\pi PR^2 \cos \theta \sin \theta d\theta.$$

This charge produces a field at the center of the sphere whose component parallel to E is

$$dE_1 = \frac{2\pi PR^2 \cos^2 \theta \sin \theta d\theta}{R^2}$$

so that the total charge on the sphere produces a field at the center equal to

$$E_1 = 2\pi P \int_0^\pi \cos^2 \theta \sin \theta d\theta = \frac{4\pi P}{3}.$$

The total electric field at the center of this sphere is then

$$E + \frac{4\pi P}{3}. \quad (13)$$

Of course, there is still the contribution to the force by the atoms inside the little sphere we have cut out, but in an isotropic medium

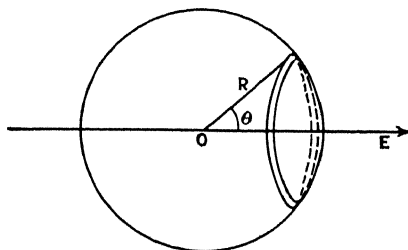


FIG. 48.—Field in spherical cavity in dielectric.

this averages zero. We can now carry over our calculations for gases if we replace E by $E + (4\pi P/3)$ in the expression for x . Thus we get

$$P = \left(E + \frac{4\pi P}{3} \right) \sum_k \frac{N_k \frac{e^2}{m}}{\omega_{0k}^2 - \omega^2 + i\omega g_k}$$

and using the relations $D = \epsilon E = E + 4\pi P$, we have

$$E + \frac{4\pi P}{3} = \frac{\epsilon + 2}{3} E$$

and we find for ϵ

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{(n - ik)^2 - 1}{(n - ik)^2 + 2} = \frac{4\pi}{3} \sum_k \frac{N_k \frac{e^2}{m}}{\omega_{0k}^2 - \omega^2 + i\omega g_k}.$$

If N represents the number of atoms, then

$$N_k = f_k N$$

and f_k gives the number of electrons of the k th kind per atom, the so called "oscillator strength," and we have

$$\frac{(n - ik)^2 - 1}{(n - ik)^2 + 2} \cdot \frac{1}{N} = \frac{4\pi}{3} \sum_k f_k \frac{e^2/m}{\omega_{0k}^2 - \omega^2 + i\omega g_k}. \quad (14)$$

In all cases of transparent substances, where we can neglect the damping force, and the index of refraction is real, we have for a given frequency of light:

$$\frac{n^2 - 1}{n^2 + 2} \cdot \frac{1}{\rho_0} = \text{constant}.$$

where ρ_0 is the density of the body, obviously proportional to N . This law, known as the Lorenz-Lorentz law, is surprisingly well obeyed for many substances. Of course, in the limit of very long electromagnetic waves, and for the electrostatic case,

$$\frac{\epsilon - 1}{\epsilon + 2} \cdot \frac{1}{\rho_0} = \text{constant},$$

giving us a relation between dielectric constant and density.

If we use the expression $E + (4\pi P/3)$ instead of E in the equation of motion of an electron, we find similarly to our equation for gases:

$$(n - ik)^2 = 1 + 4\pi \sum_k \frac{N_k e^2/m}{\bar{\omega}_k^2 - \omega^2 + i\omega g_k}, \quad (15)$$

with the only difference that instead of the natural frequency ω_{0k} of the electrons we find the apparent natural frequencies

$$\bar{\omega}_k^2 = \omega_{0k}^2 - \frac{4\pi}{3} N_k \frac{e^2}{m}. \quad (16)$$

Thus we have the same type of anomalous dispersion phenomena in solids and liquids that we have in gases.

175. Dispersion of Metals.—In metals we picture free electrons wandering about among fixed ions, and these electrons are the conduction electrons. On the average there is no resultant force on the electrons, so that under the influence of an external field we can place the force on an electron equal to eE . If we imagine the ions as rigid structures possessing no polarizability, we then have the simplest possible picture of a metal. Thus, in contrast to the bound electrons of the previous sections, we have no restoring force on these electrons. We must, however, introduce a

damping force, so that steady-state motion becomes possible. Thus we have as the equation of motion of conducting electrons:

$$m\ddot{x} + mg\dot{x} = eE. \quad (17)$$

This equation must allow an atomistic calculation of the conductivity and if the external field E is constant and in the x direction, we get as the steady-state solution of this equation

$$x = \frac{eE}{mg}t + \text{constant}.$$

Thus the velocity is $\dot{x} = eE/mg$, and if N is the number of conducting electrons per unit volume, we get for the current density

$$u = Ne\dot{x} = \frac{Ne^2E}{mg}.$$

Now by Ohm's law $u = \sigma E$, we find

$$\sigma = \frac{Ne^2}{mg} \quad (18)$$

so that we are led to an expression for conductivity from an atomic point of view. It is interesting to note that dimensionally σ and g are both of the dimensions of frequencies. We have already seen in Prob. 10, Chap. XXII, that the period associated with σ is the relaxation time, the time taken for any irregularity in charge distribution within the metal to decrease to $1/eth$ of its value, and have seen that this frequency, for good conductors, is in the ultra-violet part of the spectrum. The meaning of g is similar, as one could see by imagining an electron initially with a given velocity, and finding the time taken for its velocity to decrease to $1/eth$ of its initial value, the result being essentially the period associated with g . It seems very reasonable to suppose that approximately equal times would be required for the velocity of electrons to be damped down, and for charge irregularities to be ironed out, and, as a matter of fact, g is found to be of the same order of magnitude as σ . One can estimate g by making a guess as to the value of N , the number of free electrons per unit volume, assuming, for instance, that there is one free electron per atom, and then computing g from the equation $g = Ne^2/m\sigma$. One has, then, two independent constants characterizing the optical behavior of a metal, so that complicated results are not surprising. In addition to this, metals like other substances contain polarizable electrons, which make additional complications.

The formulas for the optical constants of a metal may be found simply by including the free or conduction electrons as a class of bound electrons whose binding force, and natural frequency, are zero. Thus

$$\begin{aligned}
 (n - ik)^2 &= 1 + \frac{4\pi N e^2/m}{-\omega^2 + i\omega g} + 4\pi \sum_k \frac{N_k e^2/m}{\omega_k^2 - \omega^2 + i\omega g_k}, \\
 n^2 - k^2 &= 1 - \frac{4\pi\sigma}{g} \frac{1}{1 + \omega^2/g^2} + \\
 &\quad 4\pi \sum_k \frac{N_k e^2}{m} \frac{\omega_k^2 - \omega^2}{(\omega_k^2 - \omega^2)^2 + (\omega g_k)^2} \\
 nk &= \frac{\sigma}{\nu} \frac{1}{1 + \omega^2/g^2} + 2\pi \sum_k \frac{N_k e^2}{m} \frac{\omega g_k}{(\omega_k^2 - \omega^2)^2 + (\omega g_k)^2}, \quad (19)
 \end{aligned}$$

where in the last two we have written Ne^2/m as σg . The summations are over the bound electrons. We notice that as the frequency becomes low compared with σ , the first term in the product nk becomes very large compared with unity, masking the effect of the bound electrons. The difference $n^2 - k^2$ does not become correspondingly large, so that in the limit, as we stated in Chap. XXII, n becomes equal to k , and both approach $\sqrt{\sigma/\nu}$, neglecting ω compared to g . It is easy to see that at low frequency $n^2 - k^2$ approaches ϵ , if in the dielectric constant we include a contribution $-4\pi\sigma/g$ from the free electrons. However, it is only at low frequencies that these simplifications enter. As the frequency enters the near infra-red or visible region, it becomes of the same magnitude as σ and g , so that the contributions of the free electrons become complicated, and at the same time nk decreases so that the contributions of the bound electrons become important. It is thus natural that experimentally the curves of $n^2 - k^2$ and nk throughout the visible part of the spectrum are very complicated, though they can be fitted fairly accurately with formulas of the type we have derived, assuming bound as well as free electrons. In the ultra-violet, the frequency becomes too high for the free electrons to follow, the contributions of the free electrons become small compared to those of the bound electrons having resonance in that region, and a metal does not behave essentially differently from an insulator.

In conclusion, we should mention that the introduction of a frictional force proportional to the velocity of the electrons is at

best an extremely rough approximation. In metals the steady state is made possible by collisions of the electrons with the ions of the lattice, and the energy of the electrons gained from the external field is thus transmitted to the lattice, excites lattice vibrations, and appears as heat, as we shall describe more in detail later. All in all, when one considers the approximate nature of this classical electron theory, it is gratifying that it checks as well as it does with experiment and assures us that a more refined atomic picture will lead to an exact theory.

Problems

1. Show that in the case of normal dispersion for the visible spectrum where there is an absorption band in the ultra-violet, the index of refraction can be written as

$$n^2 = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \dots,$$

where λ is the wave length in vacuum and A, B, C are constants.

If there is also absorption in the infra-red show that the index of refraction is then given by

$$n^2 = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \dots - A'\lambda^2 - B'\lambda^4 \dots$$

2. Measurements of H_2 gas give the following values of the index of refraction:

λ in Å.	$(n - 1)10^7$
5,462.260	1,396
4,078.991	1,426
3,342.438	1,461
2,894.452	1,499
2,535.560	1,547
2,302.870	1,594
1,935.846	1,718
1,854.637	1,760

Using the expression in Prob. 1 for n^2 in reciprocal powers of λ , calculate the best values of A, B , and C . If the measurements are made at room temperature and atmospheric pressure, calculate the resonant frequency ω_0 and wave length from these constants.

3. Prove that in the case of anomalous dispersion for gases the maximum and minimum values of n occur at the positions where the absorption coefficient reaches half its maximum value. Show that the half width of the absorption band equals the damping constant divided by the mass of an electron. Assume $g/\omega_0 < 1$.

4. For the D line of sodium the following values of the constants in the dispersion formula are found:

$$\omega_0 = 3 \times 10^{15}; g = 2 \times 10^{10}; 4\pi Ne^2/m = 10^{23}.$$

Plot the index of refraction n and the absorption coefficient k as a function of the frequency of light. Find the maximum and minimum values of the

index of refraction n . Find the maximum value of the absorption coefficient k and the half width of the absorption band in Ångström units.

5. Show that for gases the Lorenz-Lorentz law takes the approximate form $\frac{2}{3} \frac{n-1}{\rho_0} = \text{constant}$. The following measurements have been made on air (ρ_0 given in arbitrary units).

ρ_0	n
1.00	1.0002929
14.84	1.004338
42.13	1.01241
69.24	1.02044
96.16	1.02842
123.04	1.03633
149.53	1.04421
176.27	1.05213

Calculate $\frac{2}{3} \frac{n-1}{\rho_0}$ and $\frac{n^2-1}{n^2+2} \frac{1}{\rho_0}$ for each of these measurements and compare the constancy of the results (calculate to four significant figures).

6. The indices of refraction for the sodium D line, and densities in grams per cubic centimeter of some liquids at 15°C. are

	ρ_0	n
Water.....	0.9991	1.3337
Carbon bisulphide.....	1.2709	1.6320
Ethyl ether.....	0.7200	1.3558

Calculate the indices of refraction for the vapors at 0°C. and 760 mm. pressure. The observed values for the vapors are 1.000250, 1.00148, and 1.00152, respectively.

7. The quantity $\frac{m(n^2-1)}{(n^2+2)\rho_0}$ is called the refractivity of a substance if m denotes its mass. Prove that the refractivities of mixtures of substances equal the sum of the refractivities of the constituents. (Neglect damping forces from the start.)

8. Show that the molecular refractivity of a compound, defined as $\frac{n^2-1}{n^2+2} \cdot \frac{M}{\rho_0}$, where M is the molecular weight, is equal to the sum of the atomic refractivities of the atoms of which the compound is formed. (Neglect damping forces.)

9. Prove that the apparent natural frequencies $\bar{\omega}_k$, in the equation for the index of refraction for a solid or a liquid, are related to the natural frequencies ω_{k0} for the electrons in isolated atoms by the equation

$$\bar{\omega}_k^2 = \omega_{k0}^2 - \frac{4\pi}{3} \frac{N_k e^2}{m}.$$

10. For the following gases we have the following values of $(n - 1)_\infty$ extrapolated to long wave lengths:

Gas	$(n - 1)_\infty \cdot 10^6$
H ₂	136.35
N ₂	294.5
O ₂	265.3

Calculate the values of $(n - 1)_\infty$ for the following gases: H₂O, NH₃, NO, N₂O₄, O₃. The measured values are 245.6, 364.6, 288.2, 496.5, 483.6, all times 10⁶. Find the percentage discrepancy between the calculated and observed values.

CHAPTER XXV

SPHERICAL ELECTROMAGNETIC WAVES

Suppose that we have an electrical charge oscillating back and forth sinusoidally with the time. This charge will send out a spherical electromagnetic wave, radiating in all directions. There are several physical problems connected with such a wave. First, the phenomenon may be on a large scale, as in a radio antenna. Radiation from a vertical antenna, as a matter of fact, can be approximately treated by replacing the antenna by such an oscillating charge. But also on a smaller scale we can treat the radiation of short electromagnetic waves, or in other words light, from an atom which contains oscillating electrons. The electrons may have been set in motion by heat or bombardment, in which case we have the treatment of the emission of light from a luminous body; or they may be in forced motion under the action of another light wave, as in the case of the scattering of light. As a first step in the discussion of these problems, we consider spherical solutions of the wave equation, then passing on to the special case of electromagnetic fields.

176. Spherical Solutions of the Wave Equation.—The wave equation can be solved by separation in spherical coordinates, as we have seen in Probs. 6, 7, and 8 of Chap. XV, and in Sec. 130, Chap. XVIII. The solutions are of the form $e^{\pm i\omega t} \sin m\phi P_l^m(\cos \theta)R(r)$, where R satisfies a differential equation which, by a slight transformation of the results quoted above, can be written

$$\frac{d^2(rR)}{dr^2} + \left[\frac{\omega^2}{v^2} - \frac{l(l+1)}{r^2} \right] rR = 0. \quad (1)$$

The solution of the equation for R was shown in the problem quoted above to be expressible in terms of Bessel's functions, of half integral order, divided by \sqrt{r} . It proves to be possible, however, to express these functions in an alternative manner in terms of exponential or trigonometric functions, and we shall use that more elementary method in the present chapter. Further, we shall find that we have to consider only the very simplest types of spherical waves, for the purposes we are interested in.

The simplest solution in spherical coordinates is the one independent of angle, for which $l = 0$. In this case, solving Eq. (1), we have $rR = e^{\pm \frac{i\omega r}{v}}$, giving as the solution of the wave equation the functions $\frac{e^{\pm i\omega(t \pm r/v)}}{r}$, reducing to $\frac{1}{r}$ for the static case where $\omega = 0$. This represents a sinusoidal wave, traveling out along r (if we have $t - r/v$) or in along r (if we have $t + r/v$), with a velocity v , and with an amplitude which decreases as $1/r$. This decrease of amplitude is necessary if equal amounts of energy are to flow across all concentric spheres, since the intensity, proportional to the square of the amplitude, must be proportional to $1/r^2$ so that its product with the area of the sphere may be constant.

A more general spherical wave can be obtained if we are not limited to sinusoidal vibrations. Thus the wave equation in spherical coordinates, neglecting terms in θ and ϕ which are zero for solutions independent of angle, is

$$\frac{\partial^2(ru)}{\partial r^2} - \frac{1}{v^2} \frac{\partial^2(ru)}{\partial t^2} = 0, \quad (2)$$

which has a general solution $u = \frac{1}{r} \left[f\left(t - \frac{r}{v}\right) + g\left(t + \frac{r}{v}\right) \right]$, as can be proved by direct substitution, where f, g , are arbitrary functions. This represents one wave traveling out from the center, another traveling in, with arbitrary wave form, and corresponds to the solution $f\left(t - \frac{lx + my + nz}{v}\right)$ for the wave equation in rectangular coordinates, expressing a plane wave of arbitrary wave form.

More complicated waves are those which are not spherically symmetrical, but instead depend on the angles. We have seen in Sec. 140, Chap. XIX, that if $1/r$ is a solution of Laplace's equation, then $\frac{\partial}{\partial n}\left(\frac{1}{r}\right)$ is a solution, where n is an arbitrary direction. This solution represents the potential of unit dipole, the differentiation giving the difference of the potentials of two opposite charges infinitesimally close together. If θ is the angle between n and the direction in which we are finding the potential, we have $\frac{\partial}{\partial n}\left(\frac{1}{r}\right) = \frac{\partial}{\partial r}\left(\frac{1}{r}\right) \cos \theta = -\frac{1}{r^2} \cos \theta$. This is a solution

of Laplace's equation, and in terms of our standard solution in spherical coordinates it is the solution corresponding to $l = 1$, $m = 0$. The function of r is $r^{-(l+1)}$, in accordance with the results of Sec. 130. As a matter of fact, it can be shown that all the solutions of Laplace's equation, and therefore all the spherical harmonics, can be derived in this way by differentiations of the simple solution $1/r$ with respect to different directions.

In a similar way, if we are given the solution $\frac{e^{i\omega(t-r/v)}}{r}$ of the wave equation, we may differentiate with respect to n and again obtain a solution. This gives

$$u = \frac{d}{dr} \left(\frac{e^{i\omega(t-r/v)}}{r} \right) \cos \theta = \left(-\frac{1}{r^2} - \frac{i\omega}{rv} \right) e^{i\omega(t-r/v)} \cos \theta. \quad (3)$$

This is the solution corresponding to $l = 1$, as before, and the function of r in Eq. (3) is the alternative way mentioned above of writing the Bessel's function obtained by direct solution of Eq. (1). For values of r small compared with a wave length, the term $1/r^2$ is large compared with ω/rv , remembering that $\omega/v = 2\pi/\lambda$. Thus at short distances the second term in Eq. (3) can be neglected, and the function falls off as $1/r^2$, as in the static case. Further, at short distances, the quantity r/v in the exponent represents a time lag which is only a short fraction of the period of oscillation, so that we may neglect it, obtaining $-\frac{1}{r^2} \cos \theta e^{i\omega t}$, the potential we should expect from a dipole of variable moment $e^{i\omega t}$ from a quasi-stationary argument in which we supposed that the variation of the moment was so slow that we could treat the dipole instantaneously as if it were constant. On the other hand, at large distances, the other term predominates, and the solution of the wave equation falls off as $1/r$. This part of the field is called the radiation field, and we see from it that this solution for a dipole persists to large r 's just as does the spherically symmetrical solution, the intensity falling off as $1/r^2$, and the field representing a wave traveling out with velocity v . This radiation field is a characteristic feature of solutions of the wave equation, and is not present in the limiting case of Laplace's equation.

177. Scalar Potential for Oscillating Dipole.—Let a charge e oscillate up and down along the z axis, its displacement being given by the real part of $Ce^{i\omega t}$. We shall assume an equal

and opposite charge to be always at the origin, so that the whole thing is electrically neutral, and constitutes a dipole of moment $eCe^{i\omega t} = Me^{i\omega t}$. We wish to find its field. We shall do this by finding the scalar and vector potentials, first computing these directly, then in a later section showing that they can be easily obtained from another vector, called the Hertz vector. The scalar and vector potentials are solutions of D'Alembert's equations, in which the charge and current densities, respectively, appear on the right sides of the equations. These are different from zero only at the dipole, which is assumed to be of infinitesimal dimensions, so that, except at the origin, the potentials satisfy the wave equation. We must then look for solutions of the wave equation satisfying the one condition that they reduce to the correct value at the origin, or at the dipole itself. The solution (3) is a function reducing to the scalar potential of a dipole in the limiting case of a static field, and we have seen that it also reduces to the value we should expect for points close to the dipole, even in a variable field. It corresponds to the scalar potential of a unit oscillating dipole. We expect, therefore, that for the dipole of moment $Me^{i\omega t}$ the scalar potential will be

$$\phi = -M \frac{d}{dr} \left(\frac{e^{-i\omega r/c}}{r} \right) \cos \theta e^{i\omega t}, \quad (4)$$

where now we write the velocity equal to c , for the case of light waves.

178. Vector Potential.—Next we may find the vector potential, using two facts: first, $\text{div } A + (1/c)\partial\phi/\partial t = 0$; second, since the current is always along the axis, the vector potential must also be in this direction. If now A is along the z axis, we easily have $A_r = A \cos \theta$, $A_\theta = -A \sin \theta$, $A_\phi = 0$, if A is the magnitude of the vector. Let us suppose tentatively that A is a function only of r (being prepared to reject this if it does not work). Then, using the equations for the divergence in spherical coordinates, we have

$$\begin{aligned} \text{div } A &= \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 A \cos \theta) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (-A \sin^2 \theta) \\ &= \frac{dA}{dr} \cos \theta + \frac{2A}{r} \cos \theta - \frac{2A \cos \theta}{r} = \frac{dA}{dr} \cos \theta. \end{aligned}$$

Also

$$\frac{1}{c} \frac{\partial \phi}{\partial t} = \frac{i\omega \phi}{c}.$$

Hence we have

$$\frac{dA}{dr} \cos \theta - \frac{i\omega}{c} M \frac{d}{dr} \left[\frac{e^{-i\omega r/c}}{r} \right] \cos \theta e^{i\omega t} = 0.$$

This can be satisfied by

$$A = \frac{i\omega}{c} M \frac{e^{-i\omega r/c}}{r} e^{i\omega t}.$$

We note that this, which represents A_z , satisfies the wave equation, as, of course, it must. Then we have

$$\begin{aligned} A_r &= \frac{i\omega}{c} M \frac{e^{-i\omega r/c}}{r} \cos \theta e^{i\omega t}, \\ A_\theta &= -\frac{i\omega}{c} M \frac{e^{-i\omega r/c}}{r} \sin \theta e^{i\omega t}, \\ A_\phi &= 0. \end{aligned} \quad (5)$$

179. The Fields.—Let us first find the magnetic field $H = \text{curl } A$. We at once have

$$\begin{aligned} H_r = H_\theta = 0, \quad H_\phi &= \frac{1}{r} \frac{\partial}{\partial r} \left(-\frac{i\omega}{c} M e^{-i\omega r/c} \sin \theta e^{i\omega t} \right) - \frac{1}{r} \frac{\partial}{\partial \theta} \\ &\left(\frac{i\omega}{c} M \frac{e^{-i\omega r/c}}{r} \cos \theta e^{i\omega t} \right) = M \frac{\omega^2}{c^2} \frac{e^{-i\omega r/c}}{r} \sin \theta e^{i\omega t} \left(-1 - \frac{c}{i\omega r} \right). \end{aligned} \quad (6)$$

From this we see that the magnetic field always goes in circles around the axis, as we should expect from the resemblance of the problem to that of a linear current. At large distances, the second term vanishes compared with the first, leaving

$$H_\phi = -\frac{M\omega^2}{c^2} \frac{e^{-i\omega r/c}}{r} \sin \theta e^{i\omega t}. \quad (7)$$

Next we find the electric field $E = -\text{grad } \phi - \frac{1}{c} \frac{\partial A}{\partial t}$. We have

$$\begin{aligned} E_r &= \frac{\partial}{\partial r} \left[M \frac{d}{dr} \left(\frac{e^{-i\omega r/c}}{r} \right) \cos \theta e^{i\omega t} \right] + M \frac{\omega^2}{c^2} \frac{e^{-i\omega r/c}}{r} \cos \theta e^{i\omega t} \\ &= M \frac{e^{-i\omega r/c}}{r} \cos \theta e^{i\omega t} \left(\frac{2i\omega}{rc} + \frac{2}{r^2} \right), \\ E_\theta &= -\frac{1}{r} M \frac{d}{dr} \left[\frac{e^{-i\omega r/c}}{r} \right] \sin \theta e^{i\omega t} - \frac{\omega^2}{c^2} M \frac{e^{-i\omega r/c}}{r} \sin \theta e^{i\omega t} \\ &= -M \frac{\omega^2}{c^2} \frac{e^{-i\omega r/c}}{r} \sin \theta e^{i\omega t} \left(1 + \frac{c}{i\omega r} - \frac{c^2}{\omega^2 r^2} \right), \\ E_\phi &= 0. \end{aligned} \quad (8)$$

From these results we see that at large r 's (large compared with a wave length), E and H become equal to each other, at right angles, and at right angles to the direction of propagation, just as with a plane wave. They equal $M \frac{\omega^2}{c^2} \frac{e^{-i\omega r/c}}{r} \sin\theta e^{i\omega t}$, the amplitudes, apart from the sinusoidal parts, being $M \frac{\omega^2}{c^2} \frac{\sin\theta}{r}$.

On the other hand, at small distances, the electrical field approaches that calculated for the dipole by electrostatics, falling off as $1/r^3$, while the magnetic field, 90 deg. out of phase with the moment, or in phase with the current, is proportional to $1/r^2$. At intermediate distances, the transition from the one situation to the other is of a complicated form. For discussion of radiation fields, it is only the result at large r that interests us.

The law giving the electric field at large r 's can be put in an interesting form. First we take the acceleration, $-\omega^2 M e^{i\omega t}$. We imagine this to be a vector along the axis. Now if we wish the field at a certain point, we take a plane normal to r passing through this point, and project the acceleration vector on that plane, using, however, not the instantaneous value but the value at the previous time ($t - r/c$). The result, dividing by rc^2 , gives $-M \frac{\omega^2}{c^2} \frac{e^{i\omega(t-r/c)}}{r} \sin\theta$, the correct value for the field. We see

from this that the dipole sends out maximum radiation to the sides, none along the line of its motion. There is an interesting extension of this to the case of a particle vibrating, not in a line, but in an arbitrary ellipse (the most general sinusoidal motion). To get the field, we again project the acceleration vector, which is proportional to the displacement, on the normal plane. Thus the vector E in general traces out an ellipse, and the wave is elliptically polarized. An interesting case is that in which the charge rotates in a circle. Then at a point along the axis, the resulting light is circularly polarized; at a point in the plane of the circle, it is linearly polarized; between, it is elliptically polarized.

180. The Hertz Vector.—There is another interesting way of considering the dipole solution, due to Hertz. The scalar and vector potentials satisfy the relation

$$\operatorname{div} A + \frac{1}{c} \frac{\partial \phi}{\partial t} = 0.$$

It would be convenient to have only one quantity from which the electromagnetic field can be derived. It is possible to find such a quantity, a vector Π , called the Hertz vector. The above relation can be satisfied identically if we place

$$\begin{aligned} A &= \frac{1}{c} \frac{\partial \Pi}{\partial t} \\ \phi &= -\text{div } \Pi. \end{aligned} \quad (9)$$

This vector Π satisfies the wave equation with no subsidiary conditions such as are imposed on the vector and scalar potentials. If any solution Π of the wave equation is found, then this represents an electromagnetic field, and the electric and magnetic fields are given by

$$\begin{aligned} E &= \text{grad div } \Pi - \frac{1}{c^2} \frac{\partial^2 \Pi}{\partial t^2} \\ H &= \frac{1}{c} \text{curl } \frac{\partial \Pi}{\partial t}. \end{aligned} \quad (10)$$

It turns out that the Hertz vector representing the field of an oscillating dipole is simply a spherically symmetrical solution of the wave equation. The correct solution, representing an outgoing wave, is

$$\Pi = \frac{p\left(t - \frac{r}{c}\right)}{r} \quad (11)$$

so that, if p represents the dipole moment of our oscillating charge (including the time variation) pointing along the z axis, it is easy to show that the vector and scalar potentials derived from this Hertz vector are just those derived in the previous sections. For example, the vector potential

$$A = \frac{1}{cr} \frac{\partial p(t - r/c)}{\partial t}$$

and if

$$p = Me^{i\omega(t-r/c)},$$

the dipole moment,

$$\frac{\partial p}{\partial t} = i\omega M e^{i\omega(t-r/c)}$$

giving for A the value we have already found

$$A = \frac{i\omega M}{cr} e^{i\omega(t-r/c)}$$

In finding the vector potential, we must remember that Π is a vector pointing along the z axis and has the components:

$$\Pi_r = \Pi \cos \theta; \Pi_\theta = -\Pi \sin \theta; \Pi_\phi = 0.$$

If we take the divergence of this vector with a negative sign we are led to our first result for the scalar potential. The fields E and H must, of course, be the same as those discussed, since the vector and scalar potentials are identical. For convenience we write the expressions for them in vector notation. From the above equations relating E and H to Π , we have

$$E = \text{grad div} \left[\frac{p(t - r/c)}{r} \right] - \frac{1}{c^2} \frac{p''(t - r/c)}{r}$$

and

$$H = \frac{1}{c} \text{curl} \left[\frac{p'(t - r/c)}{r} \right] \quad (12)$$

where the dashes denote differentiation with respect to t . These expressions lead to the same values we have been using when p varies sinusoidally with its argument. They are somewhat more general since they hold for any periodic motion of the dipole.

181. Intensity of Radiation from a Dipole.—We can easily compute Poynting's vector, and find the total rate at which the dipole is radiating energy. Poynting's vector is evidently

$$\frac{cM^2}{4\pi} \frac{\omega^4}{c^4} \frac{\cos^2 \omega(t - r/c)}{r^2} \sin^2 \theta,$$

the time average being

$$\frac{M^2 \omega^4 \sin^2 \theta}{8\pi c^3 r^2}.$$

Let us now integrate over the surface of a sphere of radius r , to get the total radiation. The element of area is $r^2 \sin \theta d\theta d\phi$, so that the result is

$$\frac{\omega^4 M^2}{8\pi c^3} \int_0^{2\pi} \int_0^\pi \sin^3 \theta d\theta d\phi = \frac{M^2 \omega^4}{3c^3} = \frac{16\pi^4 M^2 \nu^4}{3c^3}, \quad (13)$$

if $\nu = \omega/2\pi$ is the frequency. This is a well-known formula for the radiation from a dipole. The two essential features are that the radiation is proportional to the square of the amplitude of the dipole, and to the fourth power of the frequency.

182. Scattering of Light.—In addition to direct radiation, it is important to consider the process of scattering of light. Sup-

pose that a wave, for example a plane wave, falls on a dipole of the sort we have considered. Let the dipole have an equation of motion

$$m(\ddot{x} + \omega_0^2 x) = eE,$$

if m is the mass, e the charge, of the vibrating particle, E the external field, and x the displacement. Then, letting $E = E_0 e^{i\omega t}$, we have ex , the moment, equal to

$$\frac{e^2}{m} \frac{E_0 e^{i\omega t}}{\omega_0^2 - \omega^2}.$$

This is the oscillating dipole moment produced by the field. Then the dipoles set into motion by the wave will emit light, which is scattered. The rate of emission by a single dipole is

$$\frac{\omega^4}{3c^3} \left(\frac{e^2}{m} \frac{E_0}{\omega_0^2 - \omega^2} \right)^2.$$

Often the scattering is measured by the amount of light scattered per cubic centimeter of material, divided by the intensity of the incident light. The latter is $(c/4\pi)(E \times H)$, its mean value being $cE_0^2/8\pi$. Further, the amount scattered per cubic centimeter is N times that scattered by a single dipole, if there are N dipoles and they scatter independently (as the molecules of a gas do). Hence for the scattering we have

$$\frac{8\pi N e^4}{3c^4 m^2} \frac{1}{\left[\left(\frac{\omega_0}{\omega} \right)^2 - 1 \right]^2}. \quad (14)$$

There are three important special cases of this scattering formula:

(a) *The Rayleigh Scattering Formula.*—This is what we have in the case where ω is small compared with ω_0 . Since for ordinary atoms ω_0 is a frequency in the ultra-violet, we have this condition in the visible range of the spectrum. Then we may neglect 1 compared with $(\omega_0/\omega)^2$, obtaining for the scattering

$$\frac{8\pi N e^4 \omega^4}{3c^4 m^2 \omega_0^4}. \quad (15)$$

The scattering is here proportional to ω^4 , or to $1/\lambda^4$, where λ is the wave length. This proportionality to the inverse fourth power of the wave length means that the short blue and violet waves will be scattered much more than the long red ones. An

example is the scattering of light by the sky. The air molecules scatter, and on account of the law they scatter much more blue light, resulting in the blue color. The transmitted light thus has the blue removed and looks red, explaining the color near the sun at sunset.

(b) *The Thomson Scattering Formula.*—In the other limiting case of x -rays, when the frequency is large compared with ω_0 , the scattering becomes

$$\frac{8\pi N e^4}{3c^4 m^2} \quad (16)$$

This formula gives a scattering independent of the wave length, and is very important in discussing x -ray scattering by substances.

(c) *Resonant Scattering.*—If ω is nearly equal to ω_0 , it is evident that the denominator can become very small (of course, if we consider damping, it will not vanish), resulting in a very large scattering. This phenomenon can be much more conspicuous than the two other cases. Thus a bulb filled with sodium vapor, which has a natural frequency in the visible region, illuminated with light of this color, will scatter so much light that it appears luminous. This phenomenon is called resonance scattering.

183. Polarization of Scattered Light.—We observe that, if the incident light is plane polarized, the dipoles will all vibrate along the direction of its electric vector. Thus there will be no intensity in the scattered light along this direction. The scattered light will have a maximum intensity at right angles, and it will be plane polarized. It was by experiments based on these facts that the polarization of x -rays was first found.

184. Coherence and Incoherence of Light.—In the previous paragraphs we calculated the scattering by N molecules which scatter independently by adding the intensities of the scattered radiation from each. The justification for this requires closer consideration. Since the Maxwell equations are linear the field vectors E and H satisfy the superposition principle, so that we should expect the total amplitude to be the sum of the amplitudes in the various waves, in which case the total intensity, being the square of the amplitude, would certainly not be the sum of the separate intensities. The key to this situation is found in the relations between the phases of the various waves which we are adding: if they are all in the same phase, they are said to be coherent, and the amplitudes add, while if they are in phases

having random relations to each other they are incoherent and the intensities add.

To be more precise, let us consider the sum of a number of sinusoidal waves, all of the same frequency, but of different amplitude and phase:

$$\sum_k A_k \cos(\omega t - \alpha_k) = \left(\sum_k A_k \cos \alpha_k\right) \cos \omega t + \left(\sum_k A_k \sin \alpha_k\right) \sin \omega t.$$

If all the phases should be the same, say $\alpha_k = 0$, then the amplitudes of the cosine and sine terms will be $\sum_k A_k$ and 0,

respectively, so that the amplitudes add, and the intensity is proportional to $\left(\sum_k A_k\right)^2$, or if, for instance, there are N terms of

equal amplitude, proportional to N^2 times the intensity of a single wave. On the other hand, the α 's may be completely independent of each other, meaning that each α is equally likely to have any value between 0 and 2π , independent of the others.

Then we can see that $\sum_k A_k \cos \alpha_k$ will be far less than $\sum_k A_k$, since

we shall have just about as many terms with positive values of $\cos \alpha_k$ as with negative, and the terms will just about cancel. The cancellation will not be complete, however, as we see if we compute the squares of the summations, which we must add to get the intensity. The square of the first summation, for instance, is

$$\left(\sum_k A_k \cos \alpha_k\right)^2 = \sum_k A_k^2 \cos^2 \alpha_k + \sum_{k \neq l} A_k A_l \cos \alpha_k \cos \alpha_l.$$

We must find the average of this, taking the α 's as independent. That is, we must perform the operation of integrating each α from 0 to 2π and dividing by 2π . When we do this, the terms $\cos^2 \alpha_k$ average to $\frac{1}{2}$, while the products of two independent

α 's average to zero, leaving $\frac{1}{2} \sum_k A_k^2$. The other summation

gives an equal term, so that we find that the mean square amplitude, or mean intensity, averaged over phases, is the sum of the individual intensities. This is the state of complete incoherence, in which for N waves the intensity is N times the intensity of a single wave, rather than N^2 as for the coherent case. The cancellation of waves, then, while not complete, is more and more

perfect as N increases, for N becomes a smaller and smaller fraction of N^2 as N increases.

We can now apply the idea of coherence to the scattering of light from a gas. The phase of the wave at a point P , scattered by an atom at a (Fig. 49), depends on the total path the light has traveled from the source to a , and from a to P . Since the molecules of a gas have no fixed positions with respect to each other, these paths are in a random relation to each other, the phases are incoherent, and we are justified in adding intensities. Such a procedure would not be allowed for example in discussing the scattering of x-rays by crystals, where the various atoms are in fixed lattice positions. Indeed, here we do get interference, and it is just by studying the interference patterns so obtained that

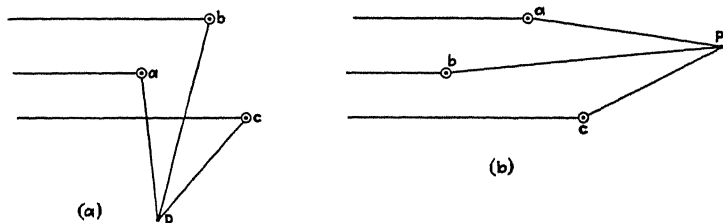


FIG. 49.—Scattering from atoms.

(a) At right angles to the incident beam, where the paths of the scattered light from the atoms a , b , c are of different and random lengths, so that there is no regular interference, and we add intensities.

(b) Scattering straight ahead, where the paths are approximately equal, and the beams interfere to produce the refracted beam.

we obtain our information about the lattice structure of crystals. Neither would the procedure be allowed in discussing the scattering from a gas in the same direction as the incident radiation, as in (b). For then the paths of the beams scattered from the various atoms are approximately equal, the waves are in phase, and they produce a resultant field at P proportional to the amplitude, rather than the intensity, of the incident wave. This scattered field can be shown to interfere with the incident wave in such a way that the resultant produces the refracted wave. The close relation of our scattering formulas to the formulas for the index of refraction, therefore, becomes clear, and it is evident that our two problems of refraction and scattering, though we have treated them separately, are really parts of the same subject. The scattering straight ahead produces refraction, and does not depend on the exact placing of the molecules. Scattering to the sides, on the other hand, does not occur unless the

molecules have a random arrangement, and then the intensity, not the amplitude, is proportional to the number of molecules.

185. Coherence and the Spectrum.—The amplitude of a wave, as a function of time, is never exactly sinusoidal, but is really a much more complicated function. It is often desirable, however, to resolve such a function into a spectrum; that is, write it as a sum of sinusoidal waves of different frequency. This can be conveniently done by Fourier series. To do this, we take a Fourier series with an extremely long period T , so long that all the phenomena we are interested in take place in a time short compared with T , so that we are not bothered by the periodicity of the series. Then, if our function is $f(t)$, we have

$$f(t) = \sum_n (A_n \cos \omega_n t + B_n \sin \omega_n t),$$

where

$$A_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos \omega_n t \, dt, \quad B_n = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin \omega_n t \, dt,$$

$$\omega_n = \frac{2\pi}{T} n. \quad (17)$$

This gives an analysis into an infinite number of sine waves, with frequencies spaced very close together (on account of the very small size of $2\pi/T$). No actual, physical wave is then perfectly sinusoidal, in the sense of having but one term in this expansion with an amplitude different from zero. We shall show in a problem that even a perfectly sinusoidal wave which persists for only a finite length of time will have appreciable amplitudes for all those frequencies within a range $\Delta\omega$, equal in order of magnitude to the reciprocal of the time during which the wave persists, so that a sine wave of long lifetime will correspond to a sharp line in the spectrum, while a rapidly interrupted wave will give a broad line. This is observed experimentally in the fact that increasing the pressure of a gas, thereby making collisions more frequent and interrupting the radiating of the atoms, broadens the spectral lines.

The intensity is proportional to $f^2(t)$, or to the square of the summation over frequencies. Just as before, this square consists of terms like $A_n^2 \cos^2 \omega_n t$, and cross terms like $A_n A_m \cos \omega_n t \cos \omega_m t$. Instantaneously none of these terms are necessarily zero. But if we average over time, the terms of the first sort average to $A_n^2/2$, while those of the second sort average to zero. The final

result, then, is that the time average intensity is the sum of the intensities of the various frequencies: $\overline{f^2(t)} = \frac{1}{2} \sum_n (A_n^2 + B_n^2)$.

We are justified in considering the terms connected with a given n to be the intensity of light of that particular frequency in the spectrum, so that we have the theoretical method of determining the spectral analysis of any disturbance. And we see that the following statement is true: on a time average, sinusoidal waves of different frequencies are always incoherent, and never interfere.

186. Coherence of Different Sources.—It is known experimentally that light from two different sources never interferes; to get interference we must take light from a single source, split it into two beams, and allow these beams to recombine. If we regarded the sources as being monochromatic, it would be hard to see why this should be, for the amplitudes of two waves of the same frequency should add, rather than the intensities, and this is the essence of interference. But when we observe that each source really is represented by a Fourier series, the situation becomes plain. For two sources are always so different that their Fourier series will be entirely different. If we analyze both of them, the phase of the radiation of frequency ω_n from one will be entirely independent of the phase of the corresponding frequency from the other. Thus if we add the disturbances, square, and average over this random relation between the phases of the two sources, the cross terms will cancel, and the intensities add. The randomness comes in this case, not in adding a great many terms of the same frequency, but in combining the terms of different frequencies, which are related in entirely independent ways in the two sources.

Problems

1. Discuss the weakening of sunlight on account of scattering, as the light passes through the atmosphere. Assume that the molecules of the atmosphere have a natural frequency at $1,800 \text{ \AA}$. (where absorption is observed). Let each molecule contain an electron of this frequency. Assume that the number of molecules is such as to give the normal barometric pressure. Find the fractional weakening of a beam due to scattering in passing through a sheet of thickness ds , and from this set up the differential equation for intensity as a function of the distance. Solve for the ratio of intensity to the intensity before striking the atmosphere, for the sun shining straight down, and for it shining at an angle of incidence of 60 deg . Constants:

$e = 4.774 \times 10^{-10}$ e.s.u., $m = 9.00 \times 10^{-28}$ gm., number of molecules in 1 gm.-mol $= 6.06 \times 10^{23}$.

2. A vibrating dipole radiates energy, and therefore its own energy decreases. Noting that the rate of radiation is proportional to the energy, set up the differential equation for the energy of the dipole as a function of the time. Find how long it takes the dipole to lose half its energy. Work out numerical values for the sort of dipole considered in Prob. 1.

3. Using the results of Prob. 2, find the equivalent damping term which would make the dipole lose energy at the same rate as the radiation. This damping is called the radiation resistance.

4. Show that the values for E and H , which we have found, satisfy Maxwell's equations, by direct calculations in polar coordinates.

5. Derive the expressions for E and H in terms of the Hertz vector Π from the equations defining Π .

6. Show that the fields E and H in terms of $p(t - r/c)$ and its time derivatives reduce to the values in terms of the dipole moment M .

7. Show that near an oscillating dipole the magnetic field is given by

$$H = -\frac{1}{cr^3}[r \times p'(t)]$$

and thus can be derived from the Biot-Savart law when we place

$$p'(t) = I(t)ds,$$

where $I(t)$ is the current and ds an element of length in the direction of the dipole.

8. Show from the Hertz vector for the dipole case, that at large distances from the dipole,

$$E = \frac{1}{c^2 r^3} \left\{ r \times \left[r \times p'' \left(t - \frac{r}{c} \right) \right] \right\}$$

and

$$H = -\frac{1}{c^2 r^2} \left[r \times p'' \left(t - \frac{r}{c} \right) \right].$$

9. Suppose we have an alternating current of maximum value I (measured in e.m.u.) in a vertical antenna of length l . Treating this as a dipole, show that the total radiation is

$$\frac{4\pi^2 c}{3} \frac{l^2 I^2}{\lambda^2}$$

Show that the equivalent resistance necessary to produce the same power loss (the radiation resistance) is

$$R = 80\pi^2 \frac{l^2}{\lambda^2}$$

if R is measured in ohms, and if we place $c = 3 \times 10^{10}$ cm. per second.

10. Find the spectrum of a disturbance which is zero up to $t = 0$, is sinusoidal until $t = T_0$, then is zero permanently. (Hint: make the period T of the Fourier series indefinitely large compared with T_0 .)

11. Find the spectrum of a disturbance which starts at $t = 0$, and is a sinusoidal damped wave after that. Show that the curve for intensity as a

function of frequency has the same form as a resonance curve, in general, and that its breadth is connected with the logarithmic decrement in the same way. This illustrates an important principle: the emission and absorption spectrum of the same substance are essentially equivalent. The resonance curve represents the absorption curve, on account of the relation of forced oscillators and dispersion, while the damped wave is the emission. (Hint: make the period T indefinitely large compared with the time taken for the oscillation to fall to $1/e$ th of its value.)

CHAPTER XXVI

HUYGENS' PRINCIPLE AND GREEN'S THEOREM

Huygens' principle is a well-known elementary method for treating the propagation of waves, and in this chapter we shall consider its mathematical background, showing its close connection with Green's theorem. The method is this: From each point of a given wave front, at $t = 0$, we assume that spherical wavelets start out. At time t , each wavelet will have a radius ct , and the envelope of these wavelets will form a new surface, which according to Huygens is simply the resulting wave front at this later time t . Thus, if the original wave front was a plane, it is easy to see that the final one will be a plane distant by the amount ct , while, if it is a sphere, the final wave front will be a concentric sphere whose radius is larger by ct . In either case this construction gives us the correct answer, agreeing with the more usual methods of computation. The one difficulty is that our construction would give a wave traveling backward, as well as one traveling forward; the solution of this difficulty appears when we use the methods of this chapter.

We may look at our process in a slightly different way, not used by Huygens, but developed later when the interference of light was being worked out. Suppose that, instead of taking the envelope of all the spherical wavelets, we consider that each of these wavelets has a certain amplitude, consisting of a sinusoidal vibration. We then add these vibrations, just as if the wavelets were being sent out by interfering sources of light, and the resulting amplitude is taken to be that in the actual wave. This process can be shown to lead to essentially the same result, and it is this which can be justified theoretically. As a further generalization, it is not necessary to take the original surface to be a wave front; it can be any surface, so long as we allow the scattered wavelets to have the suitable phase and amplitude.

Our final result, then, is this: The disturbance at a point P of a wave field may be obtained by taking an arbitrary surface,

and performing an integration over this surface. The contribution of a small element of area dS of this surface equals the amplitude at P of a spherical wave starting from dS at such a time that it reaches P at time t . That is, if the distance from dS to P is denoted by r , this wave is of the form $\frac{f(t - r/c)}{r}$.

Now the contribution, for a given wavelet, must surely be proportional to the disturbance at dS , which we may call f (a function of time and position), and to dS . Hence we have something like $\int \int \frac{f(t - r/c)}{r} dS$ for the final result. We are thus led to a formula of this sort:

$$f \text{ (at a point } P) = \text{constant} \times \int \int \frac{f(t - r/c)}{r} dS,$$

where the surface integral is over a surface surrounding P . This suggests the solution of Laplace's equation by Green's method, where we had the value of a function ϕ at an interior point of a region where $\nabla^2 \phi$ was zero as a surface integral over the boundary. As a matter of fact, an analogue to Green's theorem is the correct statement of Huygens' principle, and replaces the formula which we have derived intuitively above, and which is not just correct.

187. The Retarded Potentials.—In Chap. XXI, we have introduced scalar and vector potentials ϕ and A , giving the electric and magnetic fields by the relations

$$\begin{aligned} E &= -\text{grad } \phi - \frac{1}{c} \frac{\partial A}{\partial t} \\ H &= \text{curl } A. \end{aligned}$$

For these potentials we found the equations

$$\begin{aligned} \nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} &= -4\pi\rho \\ \nabla^2 A - \frac{1}{c^2} \frac{\partial^2 A}{\partial t^2} &= -\frac{4\pi u}{c}, \end{aligned} \quad (1)$$

or D'Alembert's equation. We ask first how to get a solution of D'Alembert's equation analogous to the simple solution

$$\phi = \int \int \int \frac{\rho}{r} dv = -\frac{1}{4\pi} \int \int \int \frac{\nabla^2 \phi}{r} dv \quad (2)$$

of Poisson's equation. We shall not carry through the proof of the solution, for that is rather complicated. But the essence of Poisson's equation is that we divide up all space into volume elements dv , and that $\rho dv/r$ is the potential of the point charge ρdv at a distance r . This potential, of course, is a solution of Laplace's equation, as is $1/r$, at all points except for $r = 0$, where the charge is located.

In a similar way, to solve D'Alembert's equation, we divide up our charge into small elements, and write the potential as the sum of the separate potentials of these small charges. The separate potentials must now be, except at $r = 0$, solutions of the wave equation. This means that, since any change of the charge will be propagated outward with the velocity c , the potential at a given point of space resulting from a particular charge cannot be derived from the instantaneous value of the charge, but must be determined, instead, by what the charge was doing at a previous instant, earlier by the time r/c required for the light to travel out from the charge to the point we are interested in. In other words, if $\rho(x, y, z, t)$ is the charge density at x, y, z at the time t , and r is the distance from x, y, z to x', y', z' , where we are finding the field, we shall expect the potential of the charge in dv to be

$$\frac{\rho(x, y, z, t - r/c)dv}{r},$$

and for the whole potential we shall have

$$\begin{aligned}\phi &= \iiint \frac{\rho(x, y, z, t - r/c)dv}{r} \\ &= -\frac{1}{4\pi} \iiint \frac{\left(\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2}\right)_{t-r/c} dv}{r}.\end{aligned}\quad (3)$$

This solution is, as a matter of fact, correct. We have already seen that $\frac{f(t - r/c)}{r}$ is a solution of the wave equation, where

f is any function, so that the integrand actually satisfies the wave equation, as in the earlier case $1/r$ satisfied Laplace's equation. The potential ϕ determined by this equation is called a retarded potential, since any change in the charge is not instantaneously observable in the potential at a distant point, but its effect is retarded on account of the finite velocity of

light. The solution for the vector potential is determined in an analogous manner.

188. Mathematical Formulation of Huygens' Principle.—In discussing the application of Green's theorem to the solution of Poisson's equations in a finite region of space, we have proved that

$$\phi = -\frac{1}{4\pi} \iiint \frac{\nabla^2 \phi}{r} dv - \frac{1}{4\pi} \iint \left(\phi \frac{\partial(1/r)}{\partial n} - \frac{1}{r} \frac{\partial \phi}{\partial n} \right) dS,$$

the result of the last paragraph being the special case where the region of integration is infinite and the surface integral drops out. We now wish to find an analogous theorem for use with D'Alembert's equation. Here again we shall not give a real derivation, for this is very complicated, but shall merely describe the formula which results, and show that it is plausible. We have already discussed the volume integral. In the surface integral, the first term gave the potential of a double layer of strength $\phi/4\pi$, the second the potential of a surface charge of magnitude $\frac{1}{4\pi} \frac{\partial \phi}{\partial n}$. Each of the terms, $\phi \frac{\partial(1/r)}{\partial n}$ and $\frac{1}{r} \frac{\partial \phi}{\partial n}$, is a solution of Laplace's equation since it represents the potential of certain charges.

In our case of the wave equation, the formula has two corresponding terms: one giving the potential of a double layer, the other of a surface charge. But now the charges change with time, so that we must use solutions of the wave equation in the integral. We have already seen that the solution of the wave equation corresponding to $\frac{1}{r}$ is $\frac{f(t-r/c)}{r}$; hence we expect the

second term to be replaced by $-\frac{1}{r} \left(\frac{\partial \phi}{\partial n} \right)_{(t-r/c)}$, where this means that the partial derivative, which is now a function of time as well as of position on the surface, is to be computed, not at t , but at $t - \frac{r}{c}$. Similarly corresponding to $\frac{\partial(1/r)}{\partial n}$, the difference of the potentials of two equal and opposite point charges at neighboring points of space, we have $\frac{\partial}{\partial n} \left\{ \frac{f(t-r/c)}{r} \right\}$. Remembering that in differentiating with respect to n we must regard r as a variable each time it occurs, this is

$$f\left(t - \frac{r}{c}\right) \frac{\partial(1/r)}{\partial n} + \frac{1}{r} \frac{\partial}{\partial n} \left[f\left(t - \frac{r}{c}\right) \right] = \\ - \frac{\cos(n, r)}{r} \left\{ \frac{f(t - r/c)}{r} + \frac{1}{c} \frac{\partial f(t - r/c)}{\partial t} \right\}.$$

where in the last term we have used the relation

$$\frac{\partial f(t - r/c)}{\partial n} = \frac{\partial f(t - r/c)}{\partial(t - r/c)} \frac{\partial(t - r/c)}{\partial n} = \frac{\partial f(t - r/c)}{\partial t} \left(-\frac{1}{c} \frac{\partial r}{\partial n} \right) \\ = -\frac{1}{c} \cos(n, r) \frac{\partial f(t - r/c)}{\partial t}.$$

We should, therefore, expect to have

$$\phi = -\frac{1}{4\pi} \iiint \frac{\left(\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} \right)_{t=r/c}}{r} dv \\ + \frac{1}{4\pi} \iint \frac{1}{r} \left\{ \left[\frac{1}{c} \left(\frac{\partial \phi}{\partial t} \right)_{t=r/c} + \frac{\phi(t - r/c)}{r} \right] \cos(n, r) \right. \\ \left. + \left(\frac{\partial \phi}{\partial n} \right)_{(t=r/c)} \right\} dS. \quad (4)$$

This, as a matter of fact, is the correct formula. The first term represents the potential due to all the charge within the volume; if there are no sources of light within this volume, the volume integral is then zero, and that is the usual case with optical applications. The surface integral represents the remaining potential as arising from a distribution of charge and double distribution about the surface, each surface element sending out a wavelet which on closer examination proves to be the Huygens' wavelet we are interested in. Thus, starting from Green's theorem and D'Alembert's equation, we have arrived at a mathematical formulation of Huygens' theorem.

To give a suggestion of the rigorous proof of this formula, we could proceed as follows: First, we notice that ϕ defined by this integral satisfies the wave equation; for since each term of the integrand separately is a solution, the sum must also be. Now it follows from this, although we have not proved it, that if the solution reduces to the correct boundary values at all points of the boundary, the solution must be the correct one, the reason being essentially that the boundary values determine a solution uniquely, so that, if we have one solution of the equation with the right boundary values, it must be the only

correct solution. We must then show that the ϕ defined by the integral actually has the correct boundary values. This could be done by a more careful treatment, and we should then have a demonstration of the formula. The more conventional proof, however, is a fairly direct though complicated application of Green's theorem.

189. Application to Optics.—We shall now take our general formula (4), and apply it to the cases we meet in optics, showing that it reduces to something like the formula which we had earlier derived intuitively. We suppose that light is emitted by a point source, and that the value of some quantity connected with, and satisfying, the wave equation (one of the components of the fields or potentials—they all satisfy the same relations) has the form

$$\frac{Ae^{2\pi i\nu(t-r_1/c)}}{r_1}, \text{ where } r_1 \text{ is the distance from the source to the point}$$

where we wish to find the disturbance. Then we wish to get the disturbance at P , not by direct calculation, but by using Huygens' principle. Suppose we take a closed surface. This surface can either surround the source, or the point P where we wish the disturbance. In any case, we have n as the normal pointing out of the part of space in which P is located. At a

point of the surface, $\phi = \frac{Ae^{2\pi i\nu(t-r_1/c)}}{r_1}$, where r_1 is the distance from the source to the point on the surface. We then have, if r is the distance from P to a point on the surface,

$$\begin{aligned} \phi\left(t - \frac{r}{c}\right) &= \frac{Ae^{2\pi i\nu[t-(r+r_1)/c]}}{r_1} \\ \frac{\partial \phi(t - r/c)}{\partial t} &= \frac{2\pi i\nu Ae^{2\pi i\nu[t-(r+r_1)/c]}}{r_1} \\ \frac{\partial \phi(t - r/c)}{\partial n} &= -A \cos(n, r_1) \left(\frac{1}{r_1} + \frac{2\pi i\nu}{c}\right) \frac{e^{2\pi i\nu[t-(r+r_1)/c]}}{r_1}, \end{aligned}$$

Thus finally

$$\begin{aligned} \phi = \frac{1}{4\pi} \int \int \frac{A}{rr_1} e^{2\pi i\nu[t-(r+r_1)/c]} &\left\{ \left(\frac{1}{r} + \frac{2\pi i\nu}{c}\right) \cos(n, r) \right. \\ &\left. - \left(\frac{1}{r_1} + \frac{2\pi i\nu}{c}\right) \cos(n, r_1) \right\} dS. \quad (5) \end{aligned}$$

In this formula, as in Chap. XXV, we have two sorts of terms, some significant at small values of r and r_1 , others at large.

We easily see that, if r and r_1 are large compared with a wave length, as is always the case in optics, the only terms we need retain are those in $\frac{2\pi i\nu}{c}$. Hence to this approximation

$$\phi = \iint \frac{iA}{2\lambda r_1 r} e^{2\pi i\nu[t-(r+r_1)/c]} [\cos(n, r) - \cos(n, r_1)] dS. \quad (6)$$

This final form suggests our earlier, intuitive formulation of Huygens' principle. The incident amplitude at dS is $\frac{A e^{2\pi i\nu(t-r_1/c)}}{r_1}$.

Now we set up, starting from dS , a wavelet whose amplitude is this value, retarded by the amount r/c , divided by r , and multiplied by the factor $\frac{i}{2\lambda} [\cos(n, r) - \cos(n, r_1)] dS$. This is just what we should expect, except for the last factor. The term i introduces a change of phase of 90 deg., not present in Huygens' form of the principle, but necessary. The term $\cos(n, r) - \cos(n, r_1)$ makes the wavelets have an amplitude which depends on angle. When r and r_1 are in opposite directions, which is the case when the surface is between the source and P , the factor approaches 2, while when r and r_1 are parallel, and the surface is beyond P , it becomes zero. This means that the wavelets do not travel backwards, thus removing the difficulty noticed earlier in Huygens' method. The wavelets have an amplitude depending on their wave length, decreasing for the longer wave lengths.

190. Integration for a Spherical Surface by Fresnel's Zones.—

Let us now carry out our integration, and verify Huygens' method, in a simple case. We take the surface to be a sphere, surrounding the source, and therefore a wave front. We note that n is the inner normal of the sphere. Thus r_1 is constant all over the sphere, and $\cos(n, r_1) = -1$ at all points, so that the formula simplifies to

$$\phi = \frac{iA e^{2\pi i\nu(t-r_1/c)}}{2\lambda r_1} \iint \frac{e^{-2\pi i r/\lambda}}{r} [\cos(n, r) + 1] dS.$$

Now suppose we introduce, as a coordinate on the sphere, the distance r from the point P ; that is, we cut the sphere with spheres concentric with P , laying off zones between them, as in Fig. 50. We can easily get the area between r and $r + dr$, and hence the element of area. Take as an axis the line joining

the source and the point P , and consider a zone making an angle between θ and $\theta + d\theta$ with the axis. The area of the zone is $2\pi r_1^2 \sin \theta d\theta$. But now by the law of cosines, if R is the distance from the source to P , $r^2 = R^2 + r_1^2 - 2Rr_1 \cos \theta$, and differentiating, $2rdr = 2Rr_1 \sin \theta d\theta$. Hence for the area of the zone we have $\frac{2\pi r r_1}{R} dr$. Introducing this, we have

$$\phi = \frac{i\pi A e^{2\pi i\nu(t-r_1/c)}}{\lambda R} \int_{r_{\min}}^{r_{\max}} e^{-2\pi i r/\lambda} [\cos(n, r) + 1] dr,$$

where $r_{\min} = R - r_1$, $r_{\max} = R + r_1$.

To carry out this integration, we use a device called Fresnel's zones, giving us an approximate value in a very elementary way.

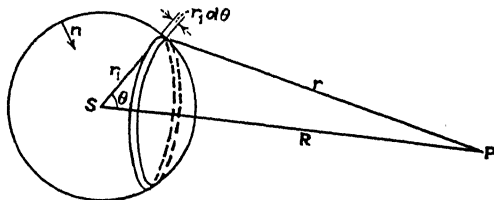


FIG. 50.—Construction for Fresnel's zones on a sphere surrounding the source.

Beginning with r_{\min} , we take a set of zones such that the outer edge of each corresponds to a value of r just half a wave length greater than the inner edge. The contributions of successive zones will almost exactly cancel. The integral, then, consists of a sum of terms, say $s_1 - s_2 + s_3 - s_4 + \dots + s_n$, where the magnitudes of s_1, s_2, \dots , vary only very slightly from one to the next. Now it is true in general that in such a series the sum is approximately half the sum of the first and last terms. We can see this as follows. We group the terms $\frac{s_1}{2} + \left(\frac{s_1}{2} - s_2 + \frac{s_3}{2}\right) + \dots +$

$\left(\frac{s_{n-2}}{2} - s_{n-1} + \frac{s_n}{2}\right) + \frac{s_n}{2}$. Now, on account of the slow varia-

tion of magnitude, we have very nearly $s_k = \frac{s_{k-1} + s_{k+1}}{2}$. If this were so, however, each of the parentheses would vanish, leaving only $\frac{s_1 + s_n}{2}$. In our case, the contribution of the first zone is to be considered, but that of the last zone is practically zero, on account of the factor $\cos(n, r) + 1$, so that the result is half the first zone.

Now, in the first zone, $\cos(n, r) + 1$ is so nearly equal to 2 that we can take it outside the integral, obtaining

$$\begin{aligned}\phi &= \frac{\pi i}{\lambda} \frac{A e^{2\pi i \nu(t-r_1/c)}}{R} \int_{R-r_1}^{R-r_1+\lambda/2} e^{-2\pi i r/\lambda} dr \\ &= -\frac{A e^{2\pi i \nu(t-r_1/c)}}{2R} (e^{-2\pi i r/\lambda}) \Big|_{R-r_1}^{R-r_1+\lambda/2} \\ &= \frac{A e^{2\pi i \nu(t-r_1/c)}}{R} e^{-2\pi i (R-r_1)/\lambda} \\ &= \frac{A e^{2\pi i \nu(t-R/c)}}{R}, \text{ the correct value.} \quad (7)\end{aligned}$$

191. The Use of Huygens' Principle.—In the derivations of this chapter we have traveled in a very roundabout way to reach a very obvious result. We naturally ask, what is Huygens' principle good for, aside from a mathematical exercise? The answer is found in the problem of diffraction. There one has certain opaque screens, with holes in them, and a light wave falling on them. If the light comes from a point source, geometrical optics would tell us that the shadow of the screen would have perfectly sharp edges. But actually this is not true; there are light and dark fringes around the edge of the shadow. If the shadow is observed at a greater and greater distance, these fringes get proportionally larger and larger, until they entirely fill the image of the hole. Finally at great distances the fringes grow in size until the resulting pattern has no resemblance at all to the geometrical image. There are then two general sorts of diffraction: first, that in which the pattern is like the geometrical image, but with diffuse edges, and which is called Fresnel diffraction; secondly, that in which the pattern is so extended that it has no resemblance to the geometrical image, and which is called Fraunhofer diffraction. Both types of diffraction, as well as the intermediate cases, can be treated by using Huygens' principle.

192. Huygens' Principle for Diffraction Problems.—Suppose that light from a point source falls on a screen containing apertures, and that we wish the amplitude at points behind the screen. Then we surround the point P , where we wish the field, by a surface consisting of the screen, and of a large surface, perhaps hemispherical, extending out beyond P , and enclosing a volume completely. We apply Huygens' principle to the surface. In doing so, we assume (1) that the amplitude of the incident wave, at points on the apertures, is the same that it would be if the

screen were absent; and (2) that immediately behind the screen, and at points of the hemispherical surface as well, the amplitude is zero, the wave being entirely cut off by the screen. This is, of course, an approximation, since at the edge of a slit, for example, the amplitude of the wave does not suddenly jump from zero to a finite value. The exact treatment is exceedingly difficult, but in the one case for which it has been worked out, it substantiates our approximations.

To find the disturbance at P , then, we integrate over the surface, but set the integrand equal to zero, except at the openings of the screen, obtaining

$$\phi = \iint \frac{iA}{2\lambda} \frac{1}{rr_1} e^{2\pi i\nu[t-(r+r_1)/c]} [\cos(n, r) - \cos(n, r_1)] dS,$$

the integral being over the openings. We note that only the edges of the openings are significant, the shape of the screen away from the opening being unimportant. Now let us assume, as is almost always true in practice, that the distances r_1 and r , from source to screen and from the screen to P , are large compared with the dimensions of the holes. Then $1/r_1$ and $[\cos(n, r) - \cos(n, r_1)]$ are so nearly constant over the aperture that we may take them outside the integral, replacing r and r_1 by mean values \bar{r} and \bar{r}_1 . If in addition we write $r + r_1$ in the exponential as $\bar{r} + \bar{r}_1 + r' + r'_1$, where r' and r'_1 are the small differences between r and r_1 and their values at some mean point of the aperture, we have finally

$$\phi = \frac{iA}{2\lambda} \frac{1}{\bar{r}\bar{r}_1} [\cos(n, \bar{r}) - \cos(n, \bar{r}_1)] e^{2\pi i\nu\left[\frac{t-(\bar{r}+\bar{r}_1)}{c}\right]} \iint e^{-2\pi i(r'+r'_1)/\lambda} dS. \quad (8)$$

The whole factor outside the integral may be taken as a constant factor so that, if we are interested only in relative intensities, we may leave it out of account. We finally have a sinusoidal vibration of which the amplitudes of the components of the two phases are proportional to $C' = \iint \cos \frac{2\pi}{\lambda} (r' + r'_1) dS$, and $S' =$

$\iint \sin \frac{2\pi}{\lambda} (r' + r'_1) dS$. Hence the intensity is proportional to $C'^2 + S'^2$, and our task is to compute this value.

193. Qualitative Discussion of Diffraction, Using Fresnel's Zones.—By using Fresnel's zones, one can see qualitatively the

explanation of the diffraction fringes, particularly in Fresnel diffraction. Suppose that we join the source S and a point P with a straight line, as in Fig. 51, and consider the point of the screen cut by this line, a point for which $r + r_1$ has a minimum value. Let us surround this point by successive closed curves in which $r + r_1$ differs from its minimum value by successive whole numbers of half wave lengths. It is not hard to see that these curves will be the intersections with the screen of a set of ellipsoids of revolution, whose foci are S and P . Hence if the line SP is approximately normal to the screen, the curves will be approximately circles. Successive zones included between successive

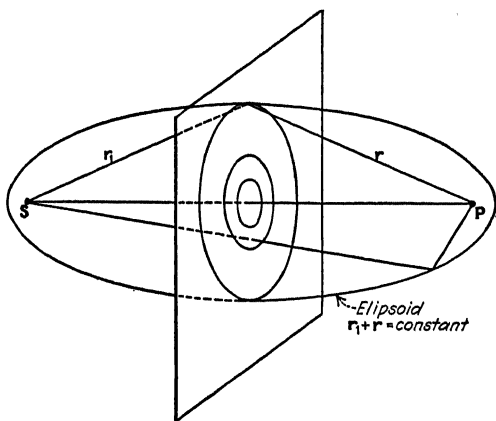


FIG. 51.—Fresnel's zones on a plane.

curves will propagate light differing by a half wave length from their neighbors. Now on the screen we may imagine the pattern of zones, and also the apertures. The whole nature of the diffraction depends on what zones are uncovered, and can transmit light, and what ones are obscured by the screen. We may distinguish three cases, shown in Fig. 52:

1. The center of the system of zones lies well inside the aperture. The central zone is entirely uncovered, as are a number of the others. As we get to larger zones, we shall come to one of which a small part is covered; then one which is more covered; and so on, until finally we come to one only slightly uncovered; and then the rest are entirely obscured. Now we can write our integral, as in paragraph 190, as a sum of integrals over the successive zones. As before, these contributions will decrease very gradually from one zone to the next. When we reach the

zones that are obscured, the decrease will become a little more rapid, but not so much as to interfere with the argument. We can still write the whole thing as half the sum of the first and the last zones. In our case, the last zone which contributes has a negligibly small area exposed, so that it contributes practically nothing, and the whole integral is half the first zone. But this gives just the intensity we should have in the absence of the screen.

2. The center of the zone system is well behind the screen (P is in the geometrical shadow). Then the first few zones are

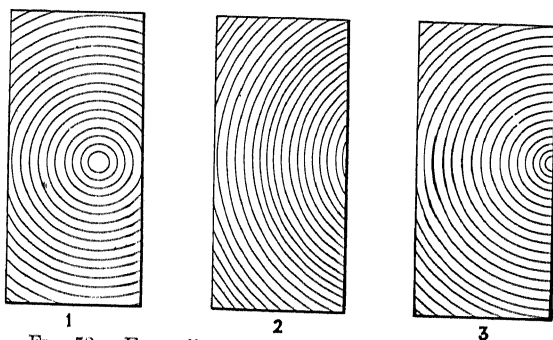


FIG. 52.—Fresnel's zones and rectangular aperture.

(1) Directly in path of light.

(2) In geometrical shadow.

(3) On edge of shadow.

obscured. A certain zone begins to be uncovered, until finally some zones are uncovered to a considerable extent. Large zones become obscured again, however. Thus in our sum, while there are terms different from zero, both the first and the last terms are zero, so that the sum is zero. The intensity well inside the geometrical shadow is zero.

3. The center of the zone system is near the edge of the screen. Then the first zone may be partly obscured, so that there is some intensity, but not so great as without the screen. Or the first zone may be entirely uncovered, but the next ones partly obscured. In these cases, the contributions from the successive zones may differ so much that our rule of taking the first and last terms is no longer correct. It is possible for the whole amplitude to be more than half the first zone, so that the intensity is actually greater than without the screen. As we move into the geometrical image from the shadow, it turns out that there is a periodic

fluctuation, on account of the uncovering of successive zones, and this explains the diffraction fringes.

Problems

1. Try to carry out exactly the integration which we did approximately by using Fresnel's zones.

2. The source is at infinity, so that a wave front is a plane. Set up Fresnel's zones, and find the breadth of the n th zone, and its area.

3. A plane wave falls on a screen in which there is a circular hole. Investigate the amplitude of the diffracted wave at a point on the axis, showing that there is alternate light and darkness as either the radius of the hole increases, or as the point moves toward or away from the screen. (Suggestion: the integral consists of a finite number of zones.)

4. A plane wave falls on a circular obstacle. Show that at a point behind the obstacle, precisely on the axis, there is illumination of the same intensity which we should have if the obstacle were not there. Explain why this would not hold for other shapes of the obstacle.

5. Take a few simple alternating series, as $1/2 - 1/3 + 1/4 - 1/5 \dots$, $1/2 - 1/4 + 1/8 \dots$, $1/2^2 - 1/3^2 + 1/4^2 - \dots$, etc., and find whether our theorem about the sum of a number of terms is verified for them. In doing this, it may be necessary to start fairly well out in the series, so as satisfy our condition that successive terms differ only slightly in magnitude.

6. Prove the statement that the boundaries of Fresnel's zones are the intersection of the screen with ellipsoids of revolution whose foci are the source and the point P . What happens to these ellipsoids as the source is removed to infinity?

CHAPTER XXVII

FRESNEL AND FRAUNHOFER DIFFRACTION

In the present chapter we proceed to the mathematical discussion of Fresnel and Fraunhofer diffraction, based on the methods of Huygens' principle derived in Chap. XXVI. The problems which we take up are Fresnel and Fraunhofer diffraction through a slit; Fraunhofer diffraction through a circular aperture; and the diffraction grating, an example of Fraunhofer diffraction. In Eq. (8) of the last chapter, we have seen that the essential step in computing the diffraction pattern is the evaluation of the integral

$$\iint e^{-2\pi i(r+r_1)/\lambda} dS,$$

where the integration is over the aperture of the screen, dS is an element of surface in the aperture, r is the distance from the source to the element dS , and r_1 the distance from the element to the point P where the field is being found. If the incident wave is a plane wave, and the plane of the aperture is a wave front, then r is the same for all elements, and the factor $e^{-2\pi ir/\lambda}$ can be cancelled out of the integral. The remaining integral, $\iint e^{-2\pi ir_1/\lambda} dS$, represents the sum at P of the amplitudes of spherical waves of equal intensity and phase starting from all points of the aperture. It is the interference of these waves which produces the diffraction pattern.

194. Comparison of Fresnel and Fraunhofer Diffraction.—

The two types of diffraction, Fresnel and Fraunhofer, arise from observing the pattern near to, or far from, the screen. Let the normal to the screen be the z axis, as in Fig. 53, and let the screen containing the aperture be at $z = 0$. The light passing through the aperture is caught on a second screen at $z = R$. Physically, the diffraction pattern has the following nature: close to the aperture, the light passes along the z axis as a column or cylinder of illumination, of cross section identical with the aperture, so that, if the screen at R is close to the aperture, the illuminated region will have the same shape as the aperture, and we speak of rectilinear propagation of the light.

As R increases, however, the column of light begins to acquire fluctuations of intensity near its boundaries, so that the pattern on the screen has fringes around the edges. This phenomenon is the Fresnel diffraction. The size of the Fresnel fringes increases proportionally to the square root of the distance R . Thus Fig. 54 shows, in its upper diagram, the slit, parallel column of light, and parabolic lines starting from the edges of the slit, indicating the position of the outer bright fringe of the Fresnel pattern, if we are sufficiently near to the slit. As R becomes larger, the fringes become so large that there are only one or two in the pattern of the aperture, and the pattern

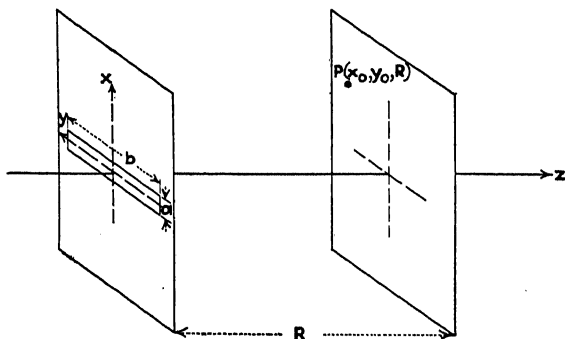


Fig. 53.—Aperture and screen for diffraction through rectangular slit.

shows but small resemblance to the shape of the aperture, though it still is of roughly the same dimensions. With further increase of R , we finally enter the region of Fraunhofer diffraction. Here the beam of light, instead of consisting of a luminous cylinder, resembles more a luminous cone (indicated by the diverging dotted lines in the top diagram of Fig. 54). Thus the Fraunhofer pattern becomes larger and larger as R increases, being in fact proportional to R , so that we can describe it by giving the angles rather than distances between different fringes. Often Fraunhofer diffraction is observed, not by placing the screen at a great distance, but by passing the light through a telescope focused on infinity. Such a telescope brings the light in a given direction to a focus at a given point of the field. Thus it separates the different Fraunhofer fringes, since each of these goes out from the source in a particular direction. In Fig. 54, diffraction patterns are shown indicating the transition from Fresnel to Fraunhofer diffraction. The pattern a illustrates the Fresnel

pattern for one edge of an infinitely wide slit. The patterns *b* to *g* represent the actual diffraction patterns from the slit, at distances indicated in the upper diagram. These patterns are all drawn to the same scale. They are drawn for a slit

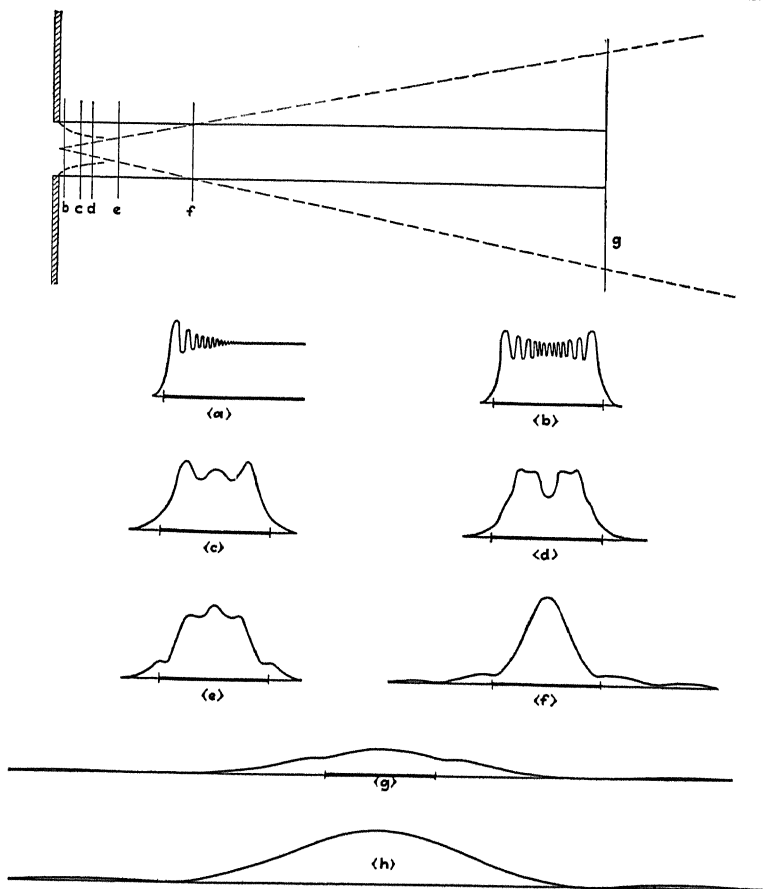


FIG. 54.—Transition from Fresnel to Fraunhofer diffraction for a slit.

(a) Fresnel pattern for edge of infinitely wide slit.

(b)–(g) Actual diffraction patterns from slit, at distances indicated in upper diagram.

(h) Fraunhofer pattern.

five wave lengths wide, for the sake of getting the figure on a diagram of reasonable scale. If the wave length were shorter, then for the same slit the distances would be stretched out to the right, and the Fraunhofer pattern would correspond to smaller angular deflections. This would be necessary to bring

the Fresnel cases far enough from the slit so that our approximations would be really applicable. Finally, in h , we give the limiting Fraunhofer pattern, not drawn to scale.

Let coordinates in the plane of the aperture be x, y , and in the plane of the screen at R let the coordinates be x_0, y_0 , as in Fig. 53. Then, if the element of area is at $x, y, 0$, and the point P at x_0, y_0, R , the distance r_1 between them is

$$r_1 = \sqrt{(x_0 - x)^2 + (y_0 - y)^2 + R^2}.$$

The integration cannot be performed with this expression for r_1 , and Fresnel and Fraunhofer diffraction lead to two different

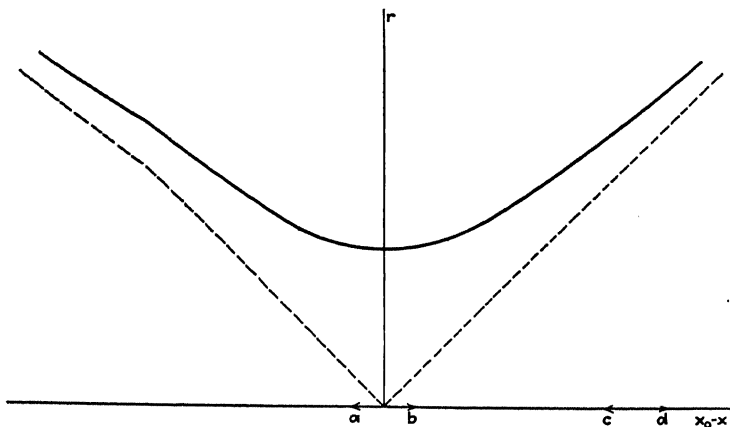


FIG. 55.— r_1 as function of $x_0 - x$: $r_1 = \sqrt{(x_0 - x)^2 + R^2}$. r_1 is the distance from a point of the aperture to a point on the screen; $x_0 - x$ is the difference between the x coordinates of the points.

approximate methods of rewriting r_1 , leading to different methods of evaluating the integral. We can see the relation of these two methods most clearly from Fig. 55, in which r_1 is plotted as a function of $x_0 - x$, for the special case where $y_0 - y = 0$. The resulting curve is a hyperbola. Now in all ordinary cases, R is large compared with the dimensions of the aperture. That is, the range of abscissas representing the dimensions of the aperture from $(x_0 - x_1)$ to $(x_0 - x_2)$, if x_1 and x_2 are the extreme coordinates of the aperture, is small compared with the distance R , the intercept of the hyperbola on the axis of ordinates. The two cases are now represented by the ranges ab and cd of abscissas, respectively. In the first, $x_0 - x_1$ and $x_0 - x_2$ are separately small, as well as their difference, and this means that the point P is almost straight behind the aperture, in the region

where the Fresnel diffraction pattern occurs. In the second, x_0 is large, of the same order of magnitude as R , showing that we are examining the pattern at a considerable angle to the normal, as we do in the Fraunhofer case. The two approximate methods can now be simply described from the curve: for Fresnel diffraction, we approximate the hyperbola near its minimum by a parabola; for Fraunhofer diffraction, we approximate it farther out by a straight line. In the first case, assuming R to be large compared with $(x_0 - x)$, we have by the binomial expansion

$$r_1 = R + \frac{1}{2} \frac{(x_0 - x)^2}{R} \dots,$$

or including the terms in y ,

$$r_1 = R + \frac{1}{2} \frac{(x_0 - x)^2 + (y_0 - y)^2}{R} + \dots \quad (1)$$

In this case, in the notation of Eq. (8) of the previous chapter, we take $\bar{r} = R$, so that r' is the remaining term of Eq. (1). For Fraunhofer diffraction, on the other hand, we have $x_0 \gg x$. Then we write $r_1^2 = (x_0^2 + y_0^2 + R^2) - 2(xx_0 + yy_0) + x^2 + y^2$, and we can neglect the terms $x^2 + y^2$. If we let $R_0^2 = x_0^2 + y_0^2 + R^2$, where R_0 measures the distance from the center of the aperture to the point P , we can use a binomial expansion, obtaining

$$r_1 = R_0 - \frac{xx_0 + yy_0}{R_0} \dots \quad (2)$$

In this case we take $\bar{r} = R_0$, so that r' is the remaining term of Eq. (2). Letting $x_0/R_0 = l$, $y_0/R_0 = m$, the direction cosines of the direction from the center of the aperture to P , we have $r' = -(lx + my) \dots$, involving the position on the screen only through the angles, so that we see at once that the pattern will travel outward radially from the aperture.

195. Fresnel Diffraction from a Slit.—Let the aperture be a slit, extending from $x = -(a/2)$ to $x = a/2$, and from $y = -(b/2)$ to $b/2$. We assume a to be small, b comparatively large, as in Fig. 53, so that it is a long narrow slit. Using the results of Eq. (1), our integral is

$$\iint e^{-2\pi i r'/\lambda} dS = \iint e^{-\pi i [(x-x_0)^2 + (y-y_0)^2]/R\lambda} dS.$$

This can be immediately factored into

$$\int_{-b/2}^{b/2} e^{-\pi i(y-y_0)^2/R\lambda} dy \int_{-a/2}^{a/2} e^{-\pi i(x-x_0)^2/R\lambda} dx.$$

Since these two integrals are of the same form, we can treat just one of them. This will prove to give fringes parallel to one set of axes. The whole pattern is then simply the combination of the two sets of fringes. The single integral, for instance the one in x , has a real part, and an imaginary part (with sign changed), equal to

$$\int_{-a/2}^{a/2} \cos \frac{\pi(x-x_0)^2}{R\lambda} dx \text{ and } \int_{-a/2}^{a/2} \sin \frac{\pi(x-x_0)^2}{R\lambda} dx. \quad (3)$$

It is customary in these integrals to make a change of variables: $\frac{(x-x_0)^2}{R\lambda} = \frac{u^2}{2}$. Then the integrals become $\sqrt{R\lambda/2}$ times C

and S , respectively, where $C = \int_{u_1}^{u_2} \cos \frac{\pi}{2} u^2 du$, $S = \int_{u_1}^{u_2} \sin \frac{\pi}{2} u^2 du$,

and where $u_1 = \frac{x_0 - a/2}{\sqrt{R\lambda/2}}$, $u_2 = \frac{x_0 + a/2}{\sqrt{R\lambda/2}}$. These integrals are called Fresnel's integrals. They cannot be explicitly evaluated, but their values have been computed by series methods.

196. Cornu's Spiral.—Let us plot the indefinite integral $\int_0^u \cos \frac{\pi}{2} u^2 du$ as abscissa, $\int_0^u \sin \frac{\pi}{2} u^2 du$ as ordinate, of a graph, as in Fig. 56. Then it is not hard to see that the resulting curve is a spiral, which is known as Cornu's spiral. To see this, we can first compute the slope. This is the differential of the ordinate, over the differential of the abscissa, or

$$\frac{\sin \frac{\pi}{2} u^2}{\cos \frac{\pi}{2} u^2} = \tan \frac{\pi}{2} u^2.$$

Thus, when u^2 increases by 4, the tangent of the curve swings around a complete cycle, and comes back to its initial value. Each point of the spiral corresponds to a particular value of u . We can show at once that the difference of u between two points is simply the length of the curve between the points. We show this for an infinitesimal element of the curve. The square of the element of length, ds^2 , is equal to the sum of the squares

of the differentials of abscissa and ordinate, or is $\cos^2\left(\frac{\pi}{2}u^2\right)du^2 + \sin^2\left(\frac{\pi}{2}u^2\right)du^2$. Hence $ds = du$, and we can integrate to get $s = u_2 - u_1$. From this fact we can make sure of the spiral nature of the curve. For one turn of the curve corresponds to an increase of u^2 by 4. That is, if u' , u'' are the values at the two

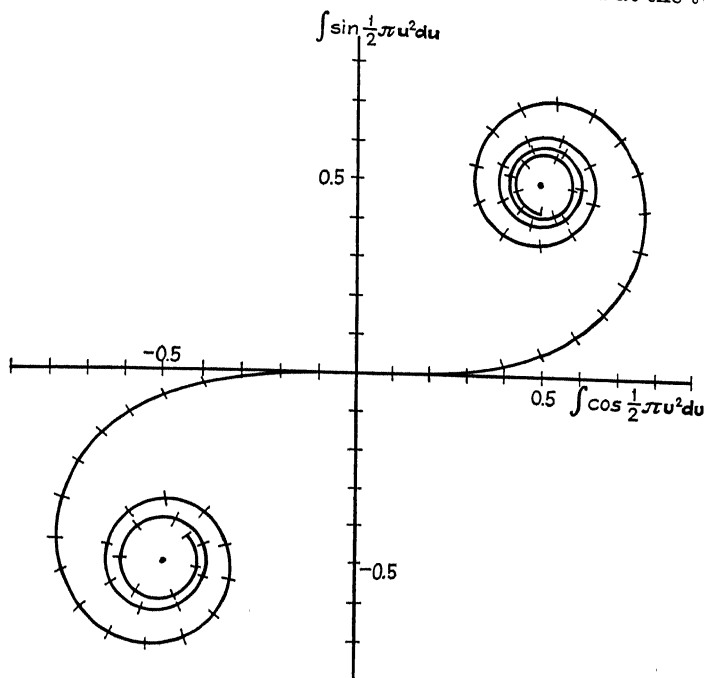


FIG. 56.—Cornu's spiral. The points of the spiral marked by cross bars correspond to increments of 0.1 unit in u .

ends, $u''^2 = u'^2 + 4$. This is $u''^2 - u'^2 = 4$, $(u'' - u')(u'' + u') = 4$, $u'' - u' = 4/(u'' + u')$. The difference $u'' - u'$ is, however, simply the length of the turn, so that we see that, as we go farther along, the turns become smaller and smaller, so that they eventually become zero, which is characteristic of a spiral. It is plain that the spiral is symmetric in the origin, having two points, for $u = \pm\infty$, for which it winds up on itself.

Let us take our spiral, mark on it the positions u_1 and u_2 corresponding to the limits of our integral, and draw the straight line connecting these points. The length of this line will then

be proportional to the amplitude of the disturbance, and its square to the intensity. This is easy to see: the horizontal component of the line is just C , and the vertical component S , so that the square of its length is $C^2 + S^2$. Knowing this, we can easily discuss the fluctuations of intensity, as seen in Fig. 54. As x_0 changes, it is plain that u_1 and u_2 increase together, their difference remaining fixed and equal to $\frac{a}{\sqrt{R\lambda/2}}$. Thus essen-

tially we have an arc of this length, sliding along the spiral, and the intensity is measured by the square of the chord between the ends of this arc. Now when x_0 is large and negative, the arc is wound up on itself, so that its ends practically meet, and the intensity is zero. This is the situation in the shadow. As x_0 approaches the value $-a/2$, however, u_2 approaches zero, so that one end of the arc has reached the center of the figure. There are two quite different cases, depending on whether $u_2 - u_1$ is large or small. If it is large (a large slit and relatively short distance R and small wave length), then u_1 will still not be unwound much at this point. The chord will then be half the value between the two end points of the spiral, and the intensity will be one-fourth its value without the screen, and will have increased uniformly in coming out of the shadow. As we go farther along the x direction, however, the arc will begin to wind up on the other half of the spiral, producing alternations of intensity at the edge of the shadow. Then for a while u_2 will be nearly at one end of the spiral, u_1 at the other, so that the intensity for some distance will be nearly constant, and the same that we should have without the slit. This is the illuminated region directly behind the slit. Finally we approach the other boundary, and u_1 commences to unwind. We then go through the same process in the opposite order. The other quite different case comes when $u_2 - u_1$ is small, which is the case for small slit, or large wave length or distance. Then there is never a time when u_1 is on one branch of the spiral and u_2 on the other. All through the central part of the pattern, therefore, there are no fluctuations of intensity. Such fluctuations come only far to one side or the other. They come about in this way: At some places in the pattern, the arc is long enough to wind up for a whole number of turns, and the chord is practically zero, while at other places it winds up for a whole number plus a half, and the chord has a maximum. The resulting fringes

are the Fraunhofer fringes which we shall now discuss by a different method.

197. Fraunhofer Diffraction from Rectangular Slit.—Using the approximation (2), our integral for Fraunhofer diffraction is $e^{-2\pi i R_0/\lambda} \iint e^{2\pi i(lx+my)/\lambda} dS$. The first term, as in Fresnel diffraction, contributes nothing to the relative intensities, and may be neglected. We then have $\iint e^{2\pi i(lx+my)/\lambda} dS$, as the integral whose absolute value measures the amplitude of the disturbance.

Let us suppose that the aperture is the same sort of rectangle considered above, extending from $-a/2$ to $a/2$ along x , from $-b/2$ to $b/2$ along y . Then the integral is

$$\begin{aligned} \int_{-a/2}^{a/2} e^{2\pi i l x/\lambda} dx \int_{-b/2}^{b/2} e^{2\pi i m y/\lambda} dy &= \frac{(e^{\pi i l a/\lambda} - e^{-\pi i l a/\lambda})}{2\pi i l/\lambda} \frac{(e^{\pi i m b/\lambda} - e^{-\pi i m b/\lambda})}{2\pi i m/\lambda} \\ &= \frac{\sin(\pi l a/\lambda)}{\pi l/\lambda} \frac{\sin(\pi m b/\lambda)}{\pi m/\lambda} \quad (4) \end{aligned}$$

The intensity is the square of this quantity. Let us consider its dependence on the position of the point P on the screen. The coordinates of this point enter only in the expressions l , m , showing that the pattern increases in size proportionally to the distance, as if it consisted of rays traveling out in straight lines from the small aperture, rather than having an approximately constant size as with the Fresnel diffraction (see Fig. 54). When we consider the detailed behavior of the intensity as a function of the angle, we find that this can be written as $\frac{a^2 \sin^2(\pi l a/\lambda)}{(\pi l a/\lambda)^2}$ times a similar function of m , giving a curve of

the form $\frac{\sin^2 \alpha}{\alpha^2}$, where $\alpha = \frac{\pi l a}{\lambda}$. This function becomes unity when $\alpha = 0$, goes to zero for $\alpha = \pi, 2\pi, 3\pi, \dots$, with maxima of intensity approximately midway between. The maxima decrease rapidly in intensity. Thus at the points $3\pi/2, 5\pi/2 \dots$ which are approximately at the second and third maxima, the intensities are only $(2/3\pi)^2, (2/5\pi)^2, \dots$ or $0.045, 0.016 \dots$, compared with the central maximum of 1. Let us see how the size of the fringes depends on the dimensions of the slit. The minima come for $\alpha = n\pi$, or $l a/\lambda = n$, $l = n\lambda/a$. Thus we see that the greater the wave length, or the smaller the dimensions of the slit, the larger the pattern becomes.

The positions of the minima can be immediately found by a very elementary argument. Assume for convenience that we are investigating the pattern at a point in the xz plane, so that $m = 0$. Then draw a plane normal to the direction l , passing through one edge of the aperture, as in Fig. 57. This represents a wave front of the diffracted wave, just as it passes one edge of the aperture. From the geometry of the system, this wave front is a distance la from the other edge, or $la/2$ from the middle of the aperture. Now, if the distance of the middle is just a whole number of half wave lengths different from the distance from the edge, the contributions of these two points to the amplitude will

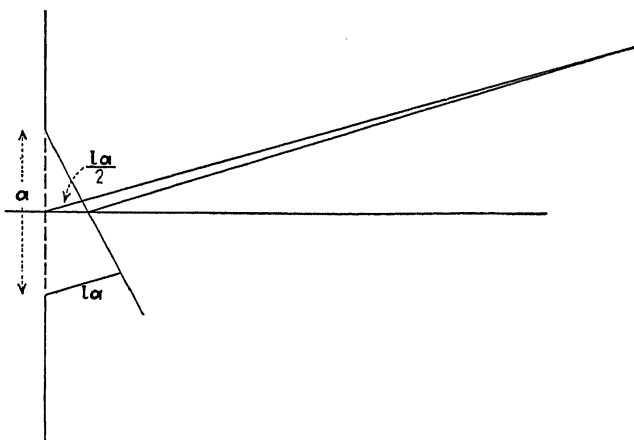


Fig. 57.—Elementary construction for Fraunhofer diffraction.

just cancel, being just out of phase. The other points of one half of the aperture can all be paired against corresponding points of the other half whose contributions are just out of phase, finally resulting in zero intensity. This situation comes about when $la/2 = n\lambda/2$, where n is an integer, or $l = n\lambda/a$, the same condition found above. Since most of the intensity falls within the first minimum, and since l is the sine of the angle between the ray and the normal to the surface, we may say that by Fraunhofer diffraction the ray is spread out through an angle λ/a .

198. The Circular Aperture.—The problem of Fraunhofer diffraction through a circular aperture is slightly more complicated mathematically. Here we must evaluate $\iint e^{2\pi i(lx + my)/\lambda} dS$ over a circle. Let us introduce polar coordinates in the plane of the aperture, so that $x = \rho \cos \theta$, $y = \rho \sin \theta$. Further, on

account of symmetry, we may take the point P to be in the xz plane, so that $m = 0$. Then if ρ_0 is the radius of the aperture, the final result is $\int_0^{2\pi} d\theta \int_0^{\rho_0} e^{2\pi i \rho \cos \theta l/\lambda} \rho d\rho$. We can integrate with respect to ρ by parts, obtaining for the integral

$$\int_0^{2\pi} d\theta \left[\frac{\rho_0 e^{2\pi i \rho_0 \cos \theta l/\lambda}}{2\pi i \cos \theta l/\lambda} - \frac{(e^{2\pi i \rho_0 \cos \theta l/\lambda} - 1)}{(2\pi i \cos \theta l/\lambda)^2} \right].$$

For the integration with respect to θ , it is necessary to expand the exponentials in series. If we do this, the integrals are in each case integrals of a power of $\cos \theta$, from 0 to 2π . These are easily evaluated, and the result, combining terms, proves to be

$$\pi \rho_0^2 \left[1 - \frac{1}{2} \left(\frac{k}{1} \right)^2 + \frac{1}{3} \left(\frac{k^2}{2!} \right)^2 - \frac{1}{4} \left(\frac{k^3}{3!} \right)^2 + \frac{1}{5} \left(\frac{k^4}{4!} \right)^2 - \dots \right],$$

where k is an abbreviation for $\pi \rho_0 l/\lambda$. If we recall the formulas for Bessel's functions, we can see without difficulty that this is equal to

$$\frac{\rho_0 \lambda}{l} J_1 \left(2\pi \rho_0 \frac{l}{\lambda} \right).$$

It is not hard, using some of the properties of Bessel's functions, to prove this formula directly, without the use of series. From the series, we see that the intensity has a maximum for $l = 0$, the center of the pattern. As l increases, we can see the behavior most easily from the expression in terms of Bessel's functions. Since J_1 has an infinite number of zeros, there are an infinite number of light and dark fringes. The first dark band comes at the first zero of J_1 , which from tables is at $2\pi \rho_0 l/\lambda = 1.2197\pi$, $\rho_0 l/\lambda = 0.61$. The next is at $\rho_0 l/\lambda = 1.16$, and so on, with maxima between. We see that, except for a numerical factor, the pattern from a circular aperture has about the same dimensions as that from a square aperture. Thus if the side of the square were equal to the diameter of the circle, $2\rho_0$, the first dark fringe would be at $2\rho_0 l/\lambda = 1$, $\rho_0 l/\lambda = 0.5$, and the next one at 1.0.

199. Resolving Power of a Lens.—Whenever light passes through a lens, it is not only refracted, but it has passed through a circular aperture, the size of the lens itself or of the diaphragm which stops it down, and as a result it is diffracted. Suppose, for example, that the lens is the objective of a telescope, and that parallel light falls on it, as from an infinitely small or distant star. Then after passing through the diaphragm, the light will no longer be a plane wave, but will have intensity in different directions, as shown in the last section. The central maximum

will have an angular diameter of $0.61 \lambda / \rho_0$, where ρ_0 is now the radius of the telescope objective. The resulting waves are just as if the light came from an object of this diameter, but passed through no diaphragm. When the telescope focuses the radiation, the result will be not a single point of light, but a circular spot surrounded by fringes, as of a star of finite diameter. For this reason, the telescope is not a perfect instrument, and one would say that its resolving power was only enough to resolve the angle $0.61 \lambda / \rho_0$. This is usually taken to mean the following: if two stars had an actual angular separation of this amount, the center of the image of one star would lie on the first dark fringe of the other, and the patterns would run into each other so that they could be just resolved. We see that the larger the aperture of the telescope, or the smaller the wave length, the better is the resolution. The same general situation holds for microscope lenses.

200. Diffraction from Several Slits; the Diffraction Grating.—

Suppose we have a number N of equal, parallel slits, equally spaced. Let each have the width a along the x axis, and let the spacing on centers be d , so that the centers come at $x = 0, d, \dots, (N-1)d$. Now let us find the Fraunhofer pattern. The part of the integral depending on y will be just as with the single slit, and we leave it out of account. We are left with

$$\int_{-a/2}^{a/2} e^{2\pi i l x / \lambda} dx + \int_{d-a/2}^{d+a/2} e^{2\pi i l x / \lambda} dx + \dots + \int_{(N-1)d-a/2}^{(N-1)d+a/2} e^{2\pi i l x / \lambda} dx.$$

But this is, as we can immediately see, simply

$$\int_{-a/2}^{a/2} e^{2\pi i l x / \lambda} dx (1 + e^{2\pi i l d / \lambda} + e^{2\pi i l 2d / \lambda} + \dots + e^{2\pi i l (N-1)d / \lambda}).$$

By the formula for the sum of a geometric series, this is $\int_{-a/2}^{a/2} e^{2\pi i l x / \lambda} dx \left(\frac{1 - e^{2\pi i l N d / \lambda}}{1 - e^{2\pi i l d / \lambda}} \right)$. Let the first term be A , the amplitude due to a single slit, which we have already evaluated. Now to find the intensity we multiply this by its conjugate, which gives

$$A^2 \frac{1 - \cos (2\pi l N d / \lambda)}{1 - \cos (2\pi l d / \lambda)} = A^2 \frac{\sin^2 (\pi l N d / \lambda)}{\sin^2 (\pi l d / \lambda)}. \quad (5)$$

That is, with N slits the actual intensity is that with one slit, but multiplied by a certain factor. This factor goes through zero when $l N d / \lambda$ is an integer, so that l equals an integer multiplied by $\lambda / N d$. This gives fringes with a narrow spacing, charac-

teristic of the whole distance Nd occupied by the set of apertures, crossing the other pattern, and they are what are usually called interference fringes, since they are due, not to diffraction from a single aperture, but to interference between different apertures. But in addition to this, the denominator results in having these fringes of different heights. The minimum height occurs when the denominator equals unity, when the fringes are of height A^2 , and the most intense fringes come when the denominator is zero. Here the ratio of numerator to denominator is evidently finite, and gives fringes of height $N^2 A^2$. Thus the greater N is, the greater the disparity in height between the largest and smallest maximum. Evidently every N th maximum will be high, and the high ones will be spaced according to the law $ld/\lambda = k$, an integer.

Now suppose N becomes very great, as in a diffraction grating. Then the small maxima will become so weak compared with the strong ones that only the latter need be considered. The latter will seem to consist of a set of sharp lines, with darkness between. These sharp lines come, as we have seen, at angles θ to the normal given by $k\lambda = d \sin \theta$, where k is an integer, and $\sin \theta = l$. This is the ordinary diffraction grating formula, where k is 0 for the central image, 1 for the first-order spectrum, 2 for the second order, etc. But we cannot entirely neglect the fact that there are other small maxima near the important ones. Thus for $ld/\lambda = k$, the intensity is $N^2 A^2$. This comes for $lNd/\lambda = Nk$. But for $lNd/\lambda = Nk + \frac{3}{2}$, we again have a secondary maximum, whose

$$\text{height is now } \frac{A^2}{\sin^2 \frac{\pi l d}{\lambda}} = \frac{A^2}{\sin^2 \pi \frac{(Nk + \frac{3}{2})}{N}} = \frac{A^2}{\sin^2 \pi \left(k + \frac{3}{2N} \right)}.$$

Now $\sin^2 \left(\pi k + \frac{3\pi}{2N} \right) = \left(\frac{3\pi}{2N} \right)^2$ approximately, if N is large,

so that the height of the maximum is $4N^2 A^2 / 9\pi^2$, or about 0.045 of the height of the highest maximum. Thus the first few secondary maxima cannot be neglected. To get an idea of the width of the region through which the intensity is considerable, we may take the width of the first maximum. From the center to the first dark fringe, this is given by the fact that at the center $lNd/\lambda = Nk$, at the dark fringe $= Nk + 1$, so that $\Delta l = \lambda/Nd$. This is closely connected with the resolving power of a grating. For a single frequency gives not a sharp set of lines, one for each order, but a set broadened by the amount we have found. Thus

two neighboring frequencies, differing by $\Delta\lambda$, could not be resolved if the first minimum of one lay opposite the maximum of the other. Since $l = \lambda k/d$, this would be the case if $\Delta l = \Delta\lambda k/d = \lambda/Nd$, or if $\Delta\lambda/\lambda = 1/Nk$. The resolving power thus increases as the number of lines in the grating increases, and as the order of the spectrum increases.

Problems

1. Carry through a discussion of Fresnel diffraction from a slit, when the source is at a finite distance, directly behind the center of the slit. In what ways will the result differ from the case we have discussed?

2. Light of wave length 6,000 Å. falls in a parallel beam on a slit 0.1 mm. broad. Work out numerical values for the intensity distribution across the slit, at three distances, first, in which the Fresnel fringes are small compared with the size of the pattern, second in which they are of the same order of magnitude, and third, in which they are Fraunhofer fringes. Either construct Cornu's spiral yourself, from tables of Fresnel's integrals, or use the one of Fig. 56.

3. Find the coordinates of the points at which Cornu's spiral winds up on itself. From the chord between these points, compute the intensity behind an infinity broad slit, which essentially means no slit at all. Find whether this agrees with what you should expect it to be.

4. Prove that the maxima of the function

$$\frac{\sin^2(\pi la/\lambda)}{(\pi la/\lambda)^2} = \frac{\sin^2 \alpha}{\alpha}$$

are determined by the equation $\alpha = \tan \alpha$. Find the first three solutions of this transcendental equation and compare them with the approximate solutions $\alpha = 3\pi/2, 5\pi/2, 7\pi/2$.

5. Discuss the Fresnel diffraction pattern caused by an edge coincident with the y axis, the screen occupying one-half the xy plane. The diffraction pattern is obtained in a plane parallel to the xy plane and a distance R from it. Plot the variation of intensity of light along the x direction from a region inside the shadow to well into the directly illuminated area. Prove that the intensity of light just at the edge of the geometrical shadow is one-fourth of its value if there were no diffraction edge.

6. Evaluate the Fresnel integrals $\int_0^u \cos \frac{\pi}{2} u^2 du$ and $\int_0^u \sin \frac{\pi}{2} u^2 du$ in a power series. What is the range of convergence of these series?

7. Evaluate the Fresnel integrals in series of the form

$$\cos^2 \frac{\pi}{2} u \Sigma_1 + \sin^2 \frac{\pi}{2} u \Sigma_2,$$

where Σ_1 and Σ_2 are power series in u . What is the range of convergence of these series?

8. Find a semiconvergent series for the Fresnel integrals of the same form as in Prob. 7 where the power series are now in inverse powers of u . (Hint: Write $\int_u^\infty \cos x^2 dx = \int_u^\infty x \cos x \frac{dx}{x}$ and integrate by parts, repeating the process.) Calculate the remainder in these series after the n th term. Show that this is smallest when n is about $x^2/2$.

CHAPTER XXVIII

WAVES, RAYS, AND WAVE MECHANICS

The beautiful success of the wave theory in explaining diffraction patterns, which we have been discussing in the last chapter, has been the best proof of the correctness of this theory. But the proof has not always gone unchallenged. Ever since the time of Newton, at least, there has been a rival theory, the corpuscular theory. Newton imagined light to consist of a stream of particles. These particles, or corpuscles, traveled in straight lines in empty space, and were reflected by mirrors as billiard balls would be by walls, making equal angles of incidence and reflection. Refraction was explained by supposing that different media had different attractions for the corpuscles. Thus glass would attract them more than air, the potential energy of a corpuscle being constant within any one medium, but being lower in glass than in air, so that the corpuscles would have a normal component of acceleration toward the glass, without corresponding tangential acceleration, and would be bent toward the normal on entering the glass. By working out this idea, the law of refraction easily follows. Newton was aware of the wave theory; Huygens was advocating it at the time. But his objection was that light travels in straight lines, whereas the waves he was familiar with, waves of sound or water waves, certainly are bent out in all directions on passing through apertures. Newton considered this to be a fatal objection to the wave theory.

The answer to this objection, of course, came later with the quantitative investigation of diffraction. In the preceding chapter, we have seen that a plane parallel wave, falling on a small aperture of dimension a , does not form a perfectly parallel ray after emerging from the hole. On the contrary, it spreads out, first by forming fringes on the edges of the ray (Fresnel diffraction), then at greater distance by developing a conical form, with definitely diverging rays (Fraunhofer diffraction). The angle of this cone is of the order of magnitude of λ/a , where λ is the

wave length. Newton was tacitly assuming that the wave length, as with sound, was large, that λ/a would be large for a small slit, and there would be large spreading out and a completely undefined ray. But it was found early in the nineteenth century that the wave length was really so small that, with apertures of ordinary size, we can neglect diffraction, and obtain an almost perfectly sharp ray, a band of light separated from the darkness by sharp, straight edges.

201. The Quantum Hypothesis.—More recently, in the present century, a more serious argument for a corpuscular theory has appeared. This is the hypothesis of quanta, originated by Planck in discussing the radiation from a heated black body. The most graphic application of this hypothesis was made by Einstein to the theory of the photoelectric effect. It is known that light of frequency ν , falling on a metal surface, liberates electrons, as for example in the photoelectric cell. Now the law of emission is remarkable: the energy of each emitted electron, independent of the intensity of the light, is a definite amount proportional to the frequency, $h\nu$, where h is Planck's constant, equal to 6.54×10^{-27} in c.g.s. units, introduced by him in his first discussion. This energy of the emitted electron is really decreased by the amount of energy it loses in penetrating the surface, so that $h\nu$ will act as a maximum energy, rather than the energy of each electron. Of course, the total emission is proportional to the intensity of the light, but increasing the intensity increases the number of electrons, not their energies.

Einstein's hypothesis to explain the photoelectric effect was that the energy of the wave was not to be computed in a continuous manner by Poynting's vector, but that it was localized in little particles or corpuscles (now called photons), each of energy $h\nu$. Then it would be perfectly obvious that if no photon fell on a spot of the metal, no electron would be ejected; but that a photon which happened to fall on a given place would transfer all its energy to an electron, being absorbed, and ceasing to exist as light. The intensity of light would be measured simply by the number of photons crossing an arbitrary surface per second, times the energy carried by each photon.

Einstein's hypothesis found many supports. One of these comes from the structure of atoms. Atoms emit monochromatic spectrum lines, falling often into regular series. Bohr was able to explain this, at least in hydrogen, the simplest atom, by assum-

ing that the atom was capable of existing only in certain definite stationary states, each of a definite energy. He supposed that radiation was not emitted continuously, as the electromagnetic field from a rotating or vibrating particle would be, but that the atom stayed in one energy level until it suddenly made a jump to a second, lower, level, with emission of a photon. If the higher energy is E_2 , the lower E_1 , the energy of the photon would be $E_2 - E_1$, so that its frequency would be $E_2/h - E_1/h$. This formula has proved to be justified by great amounts of experimental material. First, it states that the frequencies emitted by atoms should be the differences of "terms" E/h , each referring to an energy level of the atom. This is found to be true in spectroscopy, and has been the most fruitful idea in the development of that science. Even tremendously complicated spectra can now be analyzed to give a set of terms, and the number of terms is much less than the number of lines, since any pair of terms, subject to certain restrictions, gives a line. But also, Bohr was able to set up a system of mechanics to govern the hydrogen atom, very simple in its fundamentals, though different from classical mechanics, which gives a very simple formula for the energy levels, agreeing perfectly with the extremely accurate experimental values. Bohr's idea of stationary states, in turn, was tested by experiments on electron bombardment. It was found that an atom in state of energy E_1 could be bombarded by an electron. If the electron's energy, as determined from the electrical difference of potential through which it had fallen, was less than $E_2 - E_1$, where E_2 is the energy of the upper state (we consider only one), it would bounce off elastically, without loss of energy. But if its energy was $E_2 - E_1$, or greater, it would often raise the atom to the upper state, which could be proved by subsequent radiation by the atom, and would lose this amount of energy itself. This definitely verified the existence of sharp energy levels in the atom. At the same time, it furnishes an example of a very interesting phenomenon. An electron bombards an atom, loses energy $E_2 - E_1$. This energy is emitted as a photon $h\nu$. The photon falls on a metal, is absorbed, ejects a photoelectron of energy $E_2 - E_1$ (minus a little, for the work of coming through the surface). The photoelectron bombards an atom, loses its energy, which goes off as a photon. Energy, in other words, passes back and forth from electrons to photons

indiscriminately. If electrons are particles, surely photons are too.

202. The Statistical Interpretation of Wave Theory.—All these phenomena suggesting photons, and a corpuscular structure for light, must not cause one to forget that light still shows interference, and that the arguments for the wave theory are as strong as ever. Various attempts were made to set up laws of motion for the photons, which would lead to the correct laws of interference and diffraction (Newton had already done it for refraction), but without success. We can see easily why this should be so. Consider very weak light, so weak that we only have a photon every minute, for example, going through a diffraction grating. Such weak light, we know experimentally, is diffracted just like stronger light. But that means, as we saw in the last chapter, that the resolving power depends on a cooperation of the whole grating; if half of it were shut off, its resolving power would be decreased, and the intensity distribution changed. Even the single photon shows evidence of the full resolving power, in that if we make a large enough exposure to have many photons, so that we can develop the photograph and measure the blackening, which surely measures the number of photons which have struck the plate, we find the full resolving power of the grating in the final photograph. But it is difficult to imagine any law of motion of a photon which will depend on rulings over the whole face of a grating, if the photons went through only one point of it.

After such difficulties, the theory that has emerged is a combination of wave theory and corpuscular theory. It is assumed that atoms emit wave fields as in the electromagnetic theory, emitted by certain oscillators connected with the atom, and vibrating with the emitted frequencies. These waves do not carry energy, but serve merely to determine the probable motion of the photons. The rate of emission of waves by the oscillator determines the probability of emission of photons. The Poynting's vector at any point of the radiation field determines the probability that a photon will cross unit cross section normal to the radiation, per second. If the oscillator is damped with time, that indicates that the probability of emission of a photon decreases with time; that is, that the probability that the atom is in its upper, excited state, from which it could emit the radiation, is decreasing with time. One can carry such a probability connection through in detail.

Probably the most graphic picture of the probability relation between photons and waves is obtained if we imagine very weak light, in which photons come along one in several seconds, forming a diffraction pattern. The diffraction pattern is assumed to be on a screen which is capable of registering the individual photons as they come along. This screen might be a photographic plate, in which a single photon is enough to make a grain developable, or it might be a screen having slits opening into Geiger counters or other devices for registering individual photons. Of course, the only way of detecting that there was light falling on the screen would be to detect the photons. First, one photon would strike the screen, in one spot, then another photon in another spot, and so on. So long as there were only a few photons, the arrangement might seem to be haphazard. But as more and more photons were present, we could find where they were densely distributed, and where there were only a few. It would then prove to be the case that the places where photons were dense were just those places where the wave theory predicted a large intensity, and the places where there were no photons were those where the wave theory indicated darkness.

203. The Uncertainty Principle for Optics.—It is characteristic of the theory that no law of motion of photons is assumed beyond this probability; according to the present view, no such detailed laws exist. Given a plane monochromatic wave of light, we know exactly the energy of each photon ($h\nu$), and its momentum (this proves to be $h\nu/c = h/\lambda$, pointing in the direction of the wave normal), but, if the intensity is uniform over space, we have no information as to the position of the photon. If we let the plane wave fall on a slit of width a , the light passing through will be more defined as to its position in space. It will be in the form of a small ray or beam, spreading by diffraction, but still, in the region of Fresnel diffraction, of width approximately a . Thus, if x is the coordinate along the wave normal, y the coordinate at right angles, the photon will surely be in a beam whose length along the x axis is infinite, but of width only about a along the y axis, as in Fig. 58. That is, the uncertainty in the y coordinate has been reduced to a : $\Delta y = a$, if Δy is the uncertainty. At the same time, however, a compensating uncertainty in the momentum has appeared. The wave is now spreading, the wave normals making angles up to about λ/a with the x axis, as shown in Sec. 197. Thus, if the whole momentum remains $p = h/\lambda$,

this will have a component along y , equal to p times the sine of the angle between the momentum and the x axis, or approximately $p\lambda/a = h/a$. But we do not know which angle, up to the maximum, the actual deviation will make, for all we know is that the photon is somewhere in the diffraction pattern. Hence the uncertainty in y momentum is of this order of magnitude of h/a . If we call it Δp_y , we have the relation

$$\Delta y \Delta p_y = \frac{ah}{a} = h. \quad (1)$$

This is an example of the uncertainty principle, concerning the amount of uncertainty inherent in the description of the motion

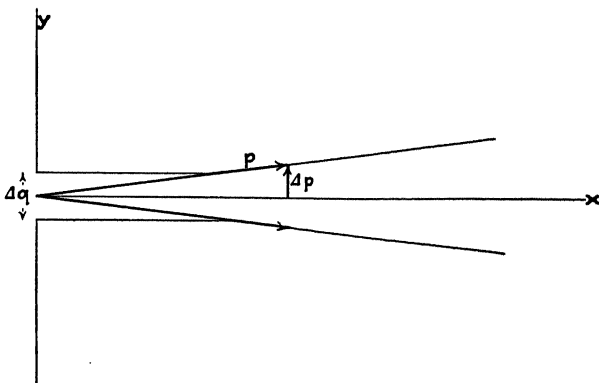


FIG. 58.—Uncertainty principle in diffraction through slit.

$$\frac{\Delta p}{p} = \frac{\lambda}{\Delta q}, \quad \Delta p \Delta q = \lambda p = h.$$

(Compare Fig. 54, top diagram).

of photons by the probability relations with wave theory. Further examination indicates that this law is very general: where a beam is limited to acquire more accurate information about the coordinates of the photon, we make a corresponding loss in our knowledge as to its momentum, and *vice versa*.

A similar relation holds between energy and time. Suppose we have a shutter over our hole, and open it and close it very rapidly, so as to allow light to pass through for only a very short interval of time Δt . Then the wave on the far side is an interrupted sinusoidal train of waves, and we know by our Fourier analysis, as in Sec. 185, that the frequency is no longer a definitely determined value, but is spread out through a frequency band of breadth $\Delta\nu$, given by $\Delta\nu/\nu = 1/(\text{number of waves in train})$.

Now the number of waves in the train is $c\Delta t$, the length of the train, divided by λ . Hence $\Delta\nu/\nu = \lambda/(c\Delta t)$, $\Delta\nu\Delta t = 1$. Using $E = h\nu$, we have

$$\Delta E\Delta t = h, \quad (2)$$

an uncertainty relation between E and t , showing that energy and time are roughly equivalent to momentum and coordinate: if we try to measure exactly when the photons go through the hole, their energy becomes slightly indeterminate. Further, here we know that the x coordinate is now determined, at any instant of time, with an accuracy $c\Delta t$: the photon must be in the little puff of light, or wave packet, sent through the pinhole while the shutter was open. Thus $\Delta x = c\Delta t$. But now the x component of momentum, which to the first order is the momentum itself, is uncertain. For $p_x = p = \frac{h\nu}{c}$, $\Delta p_x = \frac{h}{c}\Delta\nu = h/(c\Delta t) = h/\Delta x$, so that

$$\Delta x\Delta p_x = h, \quad (3)$$

again the uncertainty relation. We can, in other words, make our wave packet smaller and smaller, until it seems almost like a particle itself, and its path is the path of the photon. The wave packet will be reflected and refracted, just as large waves would be, giving the laws of motion of photons in refracting media. But if we try to go too far, making the wave packet too small, we defeat our purpose, and make it spread out by diffraction. We cannot, that is, get exactly accurate knowledge about the laws of the photon's motion from the probability relation. In some cases, this is even more obvious than here. Thus, if a wave packet is sent through a diffraction grating, it will spread out much as a plane wave would, into the various orders of the diffraction pattern. We cannot, then, make any prediction at all, except a statistical one, as to which order of the pattern a given photon will go to. We completely lose track of the paths of individual photons in a diffraction pattern.

204. Wave Mechanics.—It is now a remarkable fact that many indications point out that there is the same dualism between waves and particles in mechanics that there is in optics. We have seen one in the way energy passes from electrons to photons, and back again. We can paraphrase our earlier remark by saying that surely if photons are connected with waves, electrons are connected with waves too. But there are more substantial

reasons. In discussing the statistical relation of waves and photons, we mentioned that the electromagnetic waves were produced by oscillators, and it appears that these oscillators have only a statistical relation to the atoms. Thus we noted that the oscillators connected with radiating atoms would be exponentially damped, while the atoms were discontinuously jumping from an excited state to a lower state from which they did not radiate. This suggests a statistical connection between the oscillators and the atoms or electrons, the number of atoms in the excited state at any instant being related to the instantaneous amplitude of the corresponding oscillators, as the number of photons is related to the amplitude of the electromagnetic wave. But there are two compelling reasons which have led to the acceptance of the connection between the motion of particles and waves. The first was the experimental proof, by Davisson and Germer, G. P. Thomson, and others, that electrons can show the same sort of diffraction effects that light shows, being diffracted by crystals, and even by ruled gratings. The second was the fact, discussed by de Broglie and developed by Schrödinger, that the stationary states of atoms and molecules correspond to the various overtones of a standing wave system. Thus the waves associated with particles not only can have progressive form, connected with particles traveling along, but can also exist as standing waves, and these are precisely the oscillators which are statistically connected with the atoms, and which represent the stationary states of Bohr's theory. We shall elaborate the theory of these stationary states in succeeding chapters.

It is definitely settled, then, that mechanics is just as much a wave phenomenon as optics is. The wave mechanics leads to Newtonian mechanics as a limiting case, just as the wave theory of light leads to geometrical optics, where one treats rays only, and where one can assume that the light consists of particles following fixed paths and moving according to fixed laws. Our work, so far in this book, has been divided roughly into two sections, mechanics, and the electromagnetic theory and optics. We now commence a third section, of equivalent importance, on wave mechanics. But as the standing waves of wave mechanics are often the atoms themselves, it is natural that our treatment should be intimately bound up with the structure of matter, a subject which one can mostly leave out in

speaking of mechanics or optics, but which is of the very essence of the problem with wave mechanics.

205. Frequency and Wave Length in Wave Mechanics.—If we are considering a mechanical particle of energy E , momentum p in a given direction, we assume that associated with it is a wave (of course, not a light wave or a vibrational wave of a material medium; we are now accustomed in physics to the idea of purely mathematical waves, without reference to any medium) whose frequency ν and wave length λ are given by the equations

$$E = h\nu, \quad p = \frac{h}{\lambda}, \quad (4)$$

the wave normal being in the direction of motion of the particle. The reason why one ordinarily is not conscious of the wave nature of mechanics is the extraordinarily small wave length involved. A particle of mass 1 gm., moving with velocity 1 cm. per second, has a wave length given by $h/\lambda = mv = 1$, $\lambda = h/1 = 6.54 \times 10^{-27}$ cm., exceedingly small compared with all ordinary dimensions. If such a particle passed through a pinhole, the corresponding wave would be diffracted, but the angle of spreading would be extremely small. With other magnitudes for the mass, however, the diffraction effect can become important. Thus an electron, of mass 9×10^{-28} gm., moving, for example, with a velocity of 10^8 cm. per second, has a wave length of $\frac{h}{9 \times 10^{-28} \times 10^8} = 7.3 \times 10^{-8}$ cm., a quantity of atomic dimensions. Thus if the electron passed through an aperture of atomic size, as a hole between atoms, it could be diffracted through a large angle. It is then evident that diffraction of electrons on an atomic scale is important; in fact, we shall see in the next chapter that this is just why the atomic scale is what it is.

206. Wave Packets and the Uncertainty Principle.—Just as with light, we assume a statistical relation between the intensity of the wave and the probability of finding the particle at the corresponding point. A uniform infinite monochromatic plane wave corresponds to a particle traveling with a definite energy and momentum in a definite direction, but whose position is entirely unknown. Such a mechanical system would be approximated by electrons which had been all accelerated to the same speed in a vacuum tube, but whose individual positions we did not

know. If we wished to fix the positions, we could let the beam of electrons fall on a screen containing a pinhole. Then any electron found on the far side would have gone through the pinhole, so that we would know its y coordinate with an uncertainty Δy (using the same coordinates as with the optical case, x normal to the screen, y in the plane of the screen). After passing through, the electrons would travel practically in a straight line; but the ray will be deviated on account of diffraction, and since the law of motion of the electron is not definitely fixed, but is merely a probability law connecting it with the wave, there will be an uncertainty in its y momentum, given by $\Delta y \Delta p_y = h$. Similarly if we try to determine the x coordinate of the electron by opening and closing a shutter, so that we know exactly when it went through the hole, we thereby introduce a broadening into the spectrum of the wave, hence an uncertainty in wave length of the particle, and finally in its x component of momentum, given by $\Delta x \Delta p_x = h$. Thus the principle of uncertainty operates with particles as with photons.

The wave packet, as set up in this way, may be made extremely small without diffraction, if the wave length is as small as it often is. Thus with a particle of the mass of familiar objects, the wave function representing the motion of its center of gravity can be concentrated in a region much smaller than atomic dimensions, without being troubled by diffraction. This packet would then, in a force field, travel around in a certain way without appreciable spreading. We know at each instant that the particle is within the packet. Thus for all practical purposes the law of motion of the packet is the same as the law of motion of the particle. This then is the direction in which we look for the derivation of Newtonian mechanics from wave mechanics. We at once see that the motion of a wave packet in mechanics will be more complicated than in optics, for the wave length in mechanics, $\lambda = h/p$, changes continuously from place to place. If we have a conservative motion, for which alone it is easy to formulate wave mechanics, we have $p^2/2m + V = E$, $\lambda = h/p = h/\sqrt{2m(E - V)}$, a function of position on account of V . E stays constant, as usual, so that the frequency is constant, as in optics. But the variable λ corresponds to a variable index of refraction. There are only a few optical cases where this is true. Generally the index changes sharply from one medium to another, and the ray of light consists of segments of straight

lines. In refraction by the atmosphere, however, as in astronomy, or in the refraction by heated air over the surface of the earth, as in mirages, the path of the light rays is curved instead of sharply bent, and this corresponds to the usual mechanical case, where the paths or orbits are curved. To proceed further with the connection between wave mechanics and Newtonian mechanics, we must first investigate the shape of a ray in a case where the index changes with position. The general principle governing this is called Fermat's principle.

207. Fermat's Principle.—Assume that we have an optical system, with a ray traveling from P_1 to P_2 . We may start the ray by letting parallel light fall on a pinhole, so that really the light travels in a narrow beam, eventually reaching P_2 . We assume that the dimensions are so large that diffraction can be neglected. Then suppose we compute the time taken for light to pass from the point P_1 to P_2 along the actual ray. This

time will be $\int_{P_1}^{P_2} \frac{ds}{v}$, where the integral is a line integral, com-

puted along the ray from P_1 to P_2 , ds is the element of length along the ray, and v is the velocity, a function of position if the index of refraction changes from point to point. Next, suppose that we compute the same integral for other paths joining P_1 and P_2 , but differing in between. Since in general the integral is not independent of path, we shall get different answers. In general, if we go from one path to another, the difference of the integral between the paths will be of the same order of small quantities as the displacement of the path. But Fermat's principle says that if one path is the correct ray, and the other is slightly displaced from it, the difference in the integral is of a higher order of small quantities. This is a sort of condition met in the calculus of variations. In that subject we have

what is called the variation of an integral: $\delta \int_{P_1}^{P_2} \frac{ds}{v}$ is the variation

of the integral, and it means the difference between the integral over one path, and over another infinitely near to it. Fermat's principle says that the variation of the integral is zero for the actual path; meaning that the actual variation is infinitesimal of a higher order than the variation of path, so that it vanishes in the limit of small variation of path. The idea of the variation of an integral is closely analogous to that of the differential of a function in ordinary calculus. Thus, if the variation of an

integral is zero, for a given path, that means that the integral itself is a maximum or minimum with respect to variations of path; or, more generally, that it is stationary, not changing with

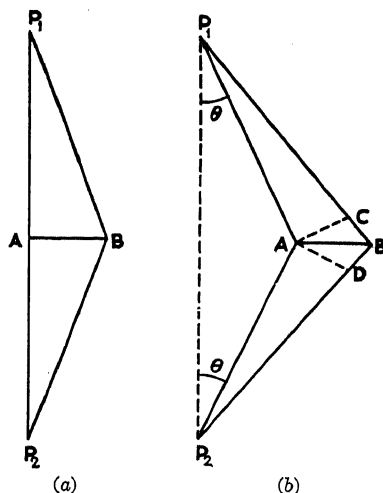


FIG. 59.—Variation of length of path. (a) The straight line P_1AP_2 differs in length from the varied path P_1BP_2 by a small quantity of the order of the square of AB .

(b) The broken line P_1AP_2 differs from P_1BP_2 by a quantity of the order of AB itself. Hence the straight line of (a), rather than the broken one of (b), is the one for which the variation of length is zero.

principle. In Fig. 59 (a), we show the straight line joining P_1 and P_2 , and also a varied path, P_1BP_2 . The length of this second path is

$$2\sqrt{(P_1A)^2 + (AB)^2} = 2(P_1A) \left[1 + \frac{1}{2} \frac{(AB)^2}{(P_1A)^2} + \dots \right] = (P_1P_2) + 2 \frac{(AB)^2}{(P_1P_2)},$$

differing from the direct path P_1P_2 by an infinitesimal of the second order, if (AB) , the deviation of the path P_1BP_2 from P_1AP_2 , is regarded as small of the first order. In other words, the path P_1AP_2 satisfies the condition that the variation of its length is zero (that is, small of the second order). On the other hand, if we started with a crooked path, as P_1AP_2 in (b), then the path P_1BP_2 differs from it approximately by the amount $(BC) + (BD)$, or approximately $2(AB) \sin \theta$, an infinitesimal

small variations of path. Setting the variation equal to zero corresponds to setting the derivative of a function equal to zero in calculus.

Let us verify Fermat's principle in two simple cases. First, we assume that v is everywhere constant, so that there are no mirrors or lenses. Then we can take v outside the integral, dividing through by it, and having $\delta \int_{P_1}^{P_2} ds = 0$. That is, the true path of light between P_1 and P_2 is that line which has minimum (or maximum) length, and joins P_1 and P_2 . Obviously the minimum is desired in this case; and the shortest line between P_1 and P_2 is a straight line, which then is the ray. Let us compute the variation of path, to check the variation

of the same order as (AB) , so that in this case the variation is not zero, and the crooked path is not the correct one.

As a second example, we take the case of reflection. In Fig. 60, consider the path P_1AP_2 , connecting P_1 and P_2 , satisfying the law of reflection on the mirror OA . This path evidently equals $P_1'AP_2$ in length, where P_1' is the image of P_1 . Similarly a slightly different path P_1BP_2 equals $P_1'BP_2$, which is therefore longer, since $P_1'AP_2$ is the straight line connecting P_1' and P_2 . In other words, P_1AP_2 makes the integral a minimum, and is the correct path. In this case we could again easily show that the integral along P_1BP_2 differed from that along P_1AP_2 by quantities in the square of AB , verifying our statement that if the path is displaced by small quantities of the first order (AB) the integral is changed only in the second order (AB^2) . A similar proof can be carried through for the case of refraction, showing that the law of refraction is given by Fermat's principle.

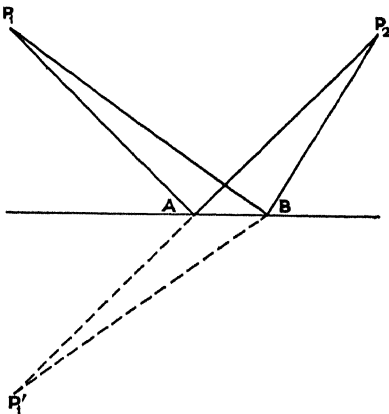


FIG. 60.—Fermat's principle for reflection. The path P_1AP_2 , equal to $P_1'AP_2$, differs in length from its neighbor P_1BP_2 by a small quantity of the order of the square of AB .

A fundamental proof of Fermat's principle can be given directly from the determination of the ray from diffraction theory. The condition that a point P_2 lie in the ray, if we discuss diffraction through the aperture by Huygens' principle as in the last chapter, is that the various paths leading from P_1 to P_2 , by going to various points of the aperture, and then being scattered in Huygens' wavelets from there to P_2 , should be approximately the same, so that the light can interfere constructively at P_2 . This means that such paths, as measured in wave lengths, are all approximately the same length. In other words, for constructive interference, $\int_{P_1}^{P_2} \frac{ds}{\lambda}$, the number of wave lengths between P_1 and P_2 ,

must be independent of slight variations in the path, or $\delta \int_{P_1}^{P_2} \frac{ds}{\lambda} = 0$.

This clearly is the condition whether λ is independent of position

or not, for, even if the waves change in length from point to point, we must still have the waves interfere to get the ray, and this still demands the same number of wave lengths along neighboring paths. Now $\lambda = v/\nu$, and since ν , the frequency, is a constant throughout the path of the light, we may then write the variation as $\nu \delta \int_{P_1}^{P_2} \frac{ds}{v} = 0$, from which, dividing by ν , we have Fermat's principle. This interpretation in terms of the interference of the waves along the ray is the fundamental meaning of Fermat's principle.

208. The Motion of Particles and the Principle of Least Action.

We shall now show that if we use the analogue to Fermat's principle in mechanics, it leads to the correct motion of the particle according to Newtonian mechanics. As we have seen, the wave problem representing the motion of a single particle whose variables we know is a ray. And the path of this ray is given by Fermat's principle, which we may write in the form $\delta \int ds/\lambda = 0$. But now in wave mechanics, $h/\lambda = p$, the momentum, so that, canceling out the constant factor h , this becomes $\delta \int p \, ds = 0$. But this is a well-known equation of ordinary mechanics: the integral $\int p \, ds$, or $\int p \, dq$, if q is the coordinate in a one-dimensional motion, is called the action, and the principle $\delta \int p \, dq = 0$, showing that the action is a maximum or more often a minimum, is called the principle of least action. And by the calculus of variations we can show that the principle of least action leads to Lagrange's equations, as the equations giving the motion of a particle which obeys the principle. This principle, or a closely related one called Hamilton's principle, also stated in terms of the calculus of variations, is often considered a fundamental formulation of the whole of mechanics, more fundamental than Newton's laws of motion, since these, in the form of Lagrange's equations, follow from it. As a matter of fact, the derivation of Lagrange's equations from the variation principle is the simplest way of deriving them, for one familiar with the calculus of variations, and leads to the equations directly in any arbitrary coordinate system. But here we have gone even farther: we have sketched the derivation of the principle of least action from wave mechanics, as the law giving the shape of a ray, determined from interference of the waves. As we see from this, wave mechanics is the fundamental branch of mechanics, and ordinary Newtonian mechanics, the mechanics of particles, is derived from it.

Problems

1. Assume in Fig. 61 that POP' is the path of the optically correct ray passing from one medium into a second one of different refractive index. Prove Fermat's principle for this case, showing that the time for the ray to pass along a slightly different path, as PAP' , differs from that along POP' by a small quantity of higher order than the distance AO . The figure is drawn so that AB , CO , are arcs of circles with centers at P and P' , respectively, and it is to be noted that for small AO , the figures AOB , AOC , are almost exactly right triangles.

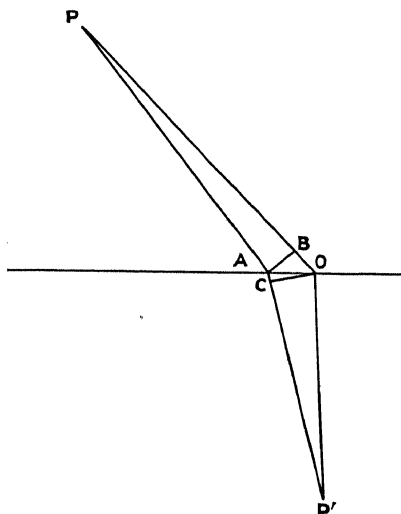


FIG. 61.—Fermat's principle for refraction.

2. An electron of charge $e = 4.774 \times 10^{-10}$ electrostatic units falls through a difference of potential of V volts (1 volt = $1/300$ e.s.u.) and bombards a target, converting all its energy into radiation, which travels out as one photon. Using the relations that the energy of the photon = $h\nu$, $\nu = c/\lambda$, where c , the velocity of light, is 3×10^{10} cm. per second, find the wave length of the resulting radiation. Find the number of volts necessary to produce visible light of wave length $5,000 \text{ \AA}$. (1 \AA . is 10^{-8} cm.); x-rays of wave length 1 \AA .; gamma rays of wave length 0.001 \AA .

3. Assume that light falls on a metal and ejects photoelectrons, the energy required to pull an electron through the surface being at least 2 volts. Find the photoelectric threshold frequency, the longest wave length which can eject electrons, remembering that the long wave lengths have small photons which have not enough energy. Discuss the effect of work function (the energy required to pull the electron out) on photoelectric threshold.

4. Newtonian mechanics becomes inaccurate when the wave length of the particle becomes of the same order of magnitude as the dimensions involved. Consider the accuracy of Newtonian mechanics in the problem of an electron

in an atom. Assume for purposes of calculation that the electron moves in a circular orbit of radius 0.5 A., with an angular momentum $h/2\pi$ (determine its speed, and hence wave length, from this fact).

5. Consider as in Prob. 4 the accuracy of Newtonian mechanics for a hydrogen atom in a hydrogen molecule. The hydrogen atom weighs about 1,800 times as much as an electron. Assume the speed of the atom to be such that its energy is the mean kinetic energy of a one-dimensional oscillator in temperature equilibrium at temperature 300° abs., or $\frac{1}{2}kT$, where $k = 1.31 \times 10^{-16}$, T is the absolute temperature. Compare the wave length with the amplitude of oscillation of the atom. To find this, assume that it oscillates with simple harmonic motion, and that its frequency of oscillation is $3,000 \text{ cm}^{-1}$. (The unit of frequency, cm^{-1} , is the frequency associated with a wave length of 1 cm.) Knowing the energy, mass, and total energy, it is then possible to find the amplitude.

6. Consider, as in Prob. 5, the same hydrogen molecule at 10° abs.; an atom of atomic weight 100, in a diatomic molecule of two like atoms, similar to the hydrogen molecule, with the same restoring force acting between the atoms (therefore with a much slower speed of vibration, on account of the larger mass), at 300° abs.; at 10° abs.

7. Consider whether the uncertainty principle is important in phenomena of astronomical magnitude. Assume a body of the mass of the earth (found from its radius of 4,000 miles, mean density 5.5), moving with a speed of 20 km. per second. Now a measurement of the position is considered, in wave mechanics, to introduce an uncertainty in the velocity, determined in terms of the uncertainty in the measurement of position by the relation $\Delta p \Delta q = h$. Suppose that the position of the body was determined in space with an error of only 1 m. (a much greater accuracy, of course, than could be really obtained). Find the corresponding uncertainty in momentum, and the angle θ through which the path is deviated by the measurement. Find how far from its original path the deviation would carry the body in a year.

8. Conjugate foci in optics are points connected by an infinite number of possible correct paths. Thus by Fermat's principle the optical path, or length of time taken to traverse the ray, is stationary for each of these paths, meaning that the optical path is the same for each. Discuss this, showing that for the conjugate foci of a simple lens the optical path is the same for each ray, carrying out the actual calculation of time.

9. Using the properties of conjugate foci mentioned in Prob. 8, prove that if a hollow ellipsoid of revolution is silvered, to form a mirror, the foci of the ellipsoid are optical conjugate foci. Prove that a paraboloidal mirror forms a perfect image of a parallel plane wave coming along its axis.

CHAPTER XXIX

SCHRÖDINGER'S EQUATION IN ONE DIMENSION

The mathematical treatment of wave mechanics starts with a wave equation, similar to those of mechanical vibrations or of light. We shall not try to derive this equation from more fundamental principles, as we derive the equation of mechanical vibration from Newton's equations, or the wave equation of optics from Maxwell's equations; there are some ways of stating wave mechanics apparently somewhat more fundamental than the wave equation, but they are not the best methods to start one's study with. We shall thus commence by postulating the wave equation, though arriving at its form by analogy with other cases. In this chapter we take only the form not involving the time, since this has a close analogy to optics. The form including the time is more remarkable, in that it involves complex quantities explicitly in its statement. We shall later treat it, separate variables in it, and show that the part independent of time is the equation treated in this chapter. This equation was first given by Schrödinger, and is called Schrödinger's equation.

As we recall, the index of refraction, and wave length, of the waves vary from point to point. This means that the differential equation is very much like that of the nonuniform string, which we discussed in Chap. XIV. We shall be able to use the same approximate solution developed for that problem. We shall also get the condition for stationary waves, corresponding to the string held at both ends. This is the so-called quantum condition, and it now determines, not the overtones of a vibrating string, but the energy levels and stationary states of atoms and other systems. The problem, as in the string, leads to expansion in orthogonal functions, and we shall consider this theory in later chapters.

209. Schrödinger's Equation.—The wave equation of optics, after the time is eliminated, can be written $\nabla^2 u + (4\pi^2/\lambda^2)u = 0$, where u is the displacement. In the mechanical problem,

$h/\lambda = p = \text{momentum}$. We assume a potential function V (wave mechanics is very difficult to formulate when there is no potential). Then the total kinetic energy is $p^2/2m$, so that $p^2/2m + V = E$, the total energy, and $p = \sqrt{2m(E - V)}$. Thus we have the equation

$$\nabla^2 u + \frac{8\pi^2 m}{h^2} (E - V) u = 0,$$

or

$$-\frac{h^2}{8\pi^2 m} \nabla^2 u + V u = E u. \quad (1)$$

These are two forms of Schrödinger's equation in the form not involving the time.

Suppose that a solution of this equation is $u(x, y, z)$. Then the corresponding solution of the problem involving the time is this times an exponential function of the time. Since the frequency ν is E/h , this is $e^{2\pi i E t/h} u(x, y, z)$. We note that the differential equation for u , and hence the resulting solution, depend on the energy E , just as the function describing the shape of a vibrating string depends on the frequency. Hence we should properly use a subscript, $u_E(x, y, z)$. The general solution would now be a sum of such solutions for all different values of E ,

$$\sum_E A_E e^{2\pi i E t/h} u_E(x, y, z), \quad (2)$$

as we had a sum of solutions as the general solution for the vibrating string.

210. One-dimensional Motion in Wave Mechanics.—For one-dimensional motion, where u is a function of x alone, Schrödinger's equation becomes

$$\frac{d^2 u}{dx^2} + \frac{8\pi^2 m}{h^2} (E - V) u = 0. \quad (3)$$

Since in general V is a function of x , this is an equation very much like that of the string with variable density but constant tension. Just as with that problem, we can easily set up an approximate solution of the problem, if the quantity $E - V$, corresponding to the density, does not change by too large a fraction of itself in a wave length, though the exact solution is generally difficult, and has been worked out in only a few special cases. The approximate solution is easily shown, by the method used in Chap. XIV, to be

$$\frac{\text{constant}}{\sqrt[4]{E - V}} e^{\pm \frac{2\pi i}{h} \int p \, dx}, \quad (4)$$

where p has the value $\sqrt{2m(E - V)}$, as before. This method of solution, as applied to wave mechanics, is often known as the Wentzel-Kramers-Brillouin method. It immediately leads to one result of physical interest, when we consider the amplitude of the wave.

We have seen in the last chapter that the intensity of the wave measured the probability of finding the particle at the corresponding point, just as in optics the intensity of the light-wave measures the probability of finding the photon. Now, if we use the wave function given above, with its complex exponential, we must evidently multiply by its conjugate to get the intensity, or the square of its amplitude:

$$\begin{aligned} u\bar{u} &= \frac{\text{constant}}{\sqrt[4]{E - V}} e^{\frac{2\pi i}{h} \int p \, dx} \times \frac{\text{constant}}{\sqrt[4]{E - V}} e^{-\frac{2\pi i}{h} \int p \, dx} \\ &= \frac{\text{constant}}{\sqrt{E - V}} = \frac{\text{constant}}{p}. \end{aligned} \quad (5)$$

To get the probability that the particle is in a small element of length ds , we must multiply through by ds , obtaining a constant $\times ds/p$. But now suppose a particle were moving along the x axis according to the Newtonian mechanics, with the same energy E , in the same potential field V . The length of time which it would spend in any small element of length ds would be ds/v , or $m \, ds/p$. Apart from the arbitrary constant, which could be determined to bring agreement, this is just like the quantum expression. If we knew that the classical particle was moving in this way, but did not know when it started, all we could say would be that the probability of finding the particle in a given region at any time was proportional to the length of time which it would have to spend in that region. In other words, our solution, of constant energy, corresponds to a classical particle whose energy is determined but whose initial time of starting is undetermined, and we can find from our wave function the probability of finding it in any region. To the approximation to which the Wentzel-Kramers-Brillouin solution is correct, the classical and quantum probabilities agree exactly, but they do not to a higher approximation. At any rate, however, we can say that the wave function is large in regions

where the particle is likely to be, or is moving slowly, and is small where the particle is moving rapidly and is unlikely to be. It should be stated that sometimes, instead of the wave function with complex exponential, we use the corresponding real wave function

$$\frac{\text{constant}}{\sqrt[4]{E - V}} \cos \frac{2\pi}{h} \int p \, dx. \quad (6)$$

In this case, the probability function has a factor of $\cos^2 \frac{2\pi}{h} \int p \, dx$, introducing a sinusoidal fluctuation of probability which must be ignored in making comparisons with the classical probability.

In the preceding paragraph, we have tacitly assumed that the kinetic energy $E - V$ was always positive, so that p was real. But in many problems, as we have seen from our discussion of classical mechanics, this is true only in limited regions, and outside these regions p becomes imaginary. Even in this case, the method of Wentzel, Kramers, and Brillouin is still formally correct. But there are two physical differences. First,

$\pm \frac{2\pi i}{h} \int p \, dx$ is now real, so that we have a real exponential, either increasing or decreasing with x , depending on the sign. Secondly, to keep the whole function real, we must make the first factor $\text{constant}/\sqrt[4]{V - E}$, which amounts to changing the constant by multiplying by $\sqrt[4]{-1}$. The approximate solution does not hold at all in the neighborhood of the point where the kinetic energy is zero, for there the wave length is infinite, and the assumption that $E - V$ changes only a little in the distance of a wave length cannot be true. But we can easily see how to construct an approximate solution in this region, for the differential equation here is simply $d^2u/dx^2 = 0$, the equation of a straight line; the actual curve of u against x , as we readily see, has a point of inflection at the point where $E = V$, being concave downward where the kinetic energy is positive, concave upward where it is negative. We can then take the exponential solution in the region of negative kinetic energy, and the oscillatory one in the region of positive kinetic energy, and join them by a line which is approximately straight. It is obvious, as we see for instance in Fig. 62, that, if we know beforehand the constants of the exponential solution (as for instance the amplitudes of the two terms, one increasing and the other decreasing exponentially, which we must add to get the

complete solution) the initial value and slope of the sinusoidal solution must be definitely determined to make the two join smoothly. That is, the phase of the sinusoidal solution, or the amplitudes of sine and cosine functions which we add together,

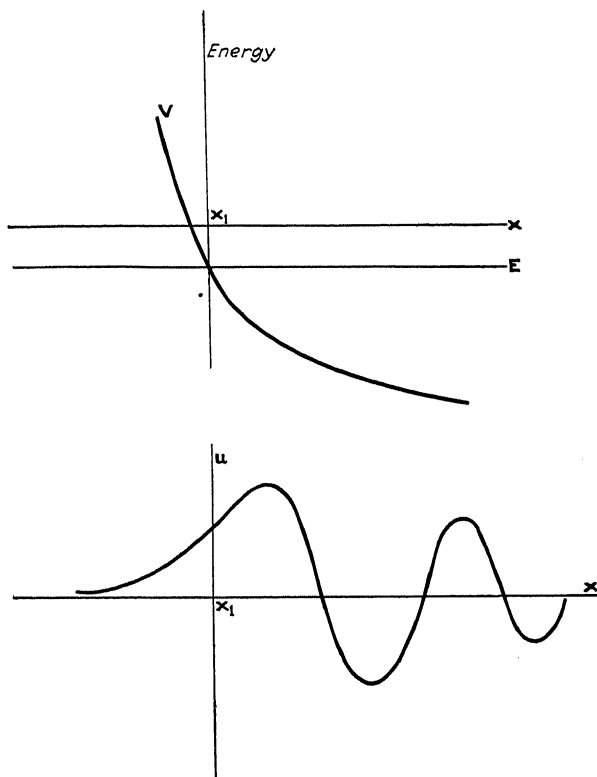


FIG. 62.—Joining of exponential and sinusoidal functions at point where $p = 0$. Upper curve shows potential and total energy against x , lower curve shows wave function. The exponential part of the function is so chosen that the amplitude of the term increasing exponentially with decreasing x is zero; otherwise the function would go to infinity instead of asymptotically to zero as x became negatively infinite.

are determined. The same thing is true at every such boundary that we cross; if we once determine the two arbitrary constants in one part of the region, the whole function is determined, to make exponential and sinusoidal curves join smoothly. This must naturally be true, since the differential equation is one of the second order, with just two arbitrary constants.

211. Boundary Conditions in One-dimensional Motion.—

Suppose, first, that we consider a mechanical problem where the kinetic energy is always positive. Then there are no regions where the wave function is exponential; it is always sinusoidal, of finite amplitude. For any energy we have two solutions, which, bringing in the time but writing in exponential form, are

$$\frac{\text{constant}}{\sqrt[4]{E - V}} e^{\frac{2\pi i}{h}(Et \pm \int p \, dx)}, \quad (7)$$

of which the real parts represent progressive waves traveling to left or right along the x axis. This corresponds to the fact that the corresponding mechanical particles can travel in either direction, and, as we have seen, the intensity of the wave at any point properly agrees with the probability that the particle should be in that region, as computed classically on the assumption that we do not know when the particle started.

Next let us assume that $E - V$ remains positive to infinity in one direction, say to the right, but becomes negative to the left of a certain point, say $x = x_1$, as in Fig. 62. The solution will then be exponential to the left of $x = x_1$. But in general it will be a linear combination of two exponential functions, one increasing exponentially in magnitude to ∞ as x approaches $-\infty$, the other decreasing exponentially to zero. If the amplitude of the former is different from zero, then the intensity of the wave will be infinite at $-\infty$, meaning that the probability of finding the particle at $-\infty$ is infinitely greater than of finding it anywhere else. This is ordinarily not the physical situation we wish to describe; hence we must assume that the amplitude is zero, and that the solution to the left of x_1 has just the one term

$$\frac{A}{\sqrt[4]{V - E}} e^{\frac{2\pi}{h} \int^x \sqrt{2m(V - E)} \, dx}, \quad (8)$$

which goes to zero at $x = -\infty$. But now this comes up to the point $x = x_1$ with a definitely determined slope (or rather, the ratio of slope to function, in which the arbitrary constant factor cancels out, is definitely determined). Then there is just a very definite sinusoidal function which joins onto this, as Fig. 62 suggests: the approximate solution for $x > x_1$ is given by

$$\frac{B}{\sqrt[4]{E - V}} \sin \left(\frac{2\pi}{h} \int_{x_1}^x p \, dx + \alpha \right). \quad (9)$$

It can be shown that for continuity of Eqs. (8) and (9) at x_1 we must have $\alpha = \pi/4 = 45^\circ$. This statement means that the sine curve, instead of having a node at x_1 , has already at that point passed through an eighth of a wave length. It is as if this eighth wave length were stretched out to infinity to form the exponential part of the curve.

We have seen, then, that a boundary where $E = V$ imposes a definite boundary condition on the solution. In our problem where the motion extends to infinity in one direction, the condition can be always satisfied, by proper choice of phase and amplitude of the sinusoidal function, as we have seen. But there are two interesting results of our calculation. First, the wave in the region where kinetic energy is positive becomes now a real function of position, or correspondingly a real function of time. In other words, it is a standing wave, not a progressive wave. It corresponds to superposed progressive waves traveling with equal intensity in both directions. The progressive wave approaching from the right is reflected at the boundary, and turns back without diminution of intensity on the reflection. The mechanical situation is that the particle, approaching the point where kinetic energy is negative, is reflected and turned back, just as it would be in the same problem in classical mechanics. But the other interesting result is that, on account of the exponential terms to the left of $x = 0$, the particles can slightly penetrate the region where kinetic energy is negative. On account of the rapid dying out of the exponential, this effect is not large, but we shall see in the next section that there can be cases where it is very important physically. This penetration by an exponential wave has an analogy in optics: a wave of light approaching an optically rarer medium at an angle greater than the critical angle is totally reflected, but at the same time, as we have seen in Sec. 168, Chap. XXIII, there is a disturbance, dying out exponentially, in the rarer medium, almost exactly equivalent to what we have here.

212. The Penetration of Barriers.—The exponential penetration of particles into the region of negative kinetic energy has as a result that in wave mechanics, unlike classical mechanics, a particle can go from one region of positive kinetic energy to another, even though there is a barrier of negative kinetic energy between. Such barriers are found, for example, in some cases at the surface of a metal, where the electrons in emerging from

the metal, for example at high temperature in thermionic emission, may find a surface layer of atoms, exerting on them such a strong repulsive force that they would be unable to penetrate on classical mechanics, but can in quantum theory. Suppose that we have a simple barrier of the sort shown in Fig. 63, where the potential has one constant value to the left of x_0 , a second high value between x_0 and x_1 , and a third lower value to the right of x_1 , and where $E - V$ is negative only between x_0 and x_1 . The corresponding problem in metals is that where the region to the left of x_0 represents the interior of the metal, that to the right of x_1 the space outside, that between x_0 and x_1 the surface layer or

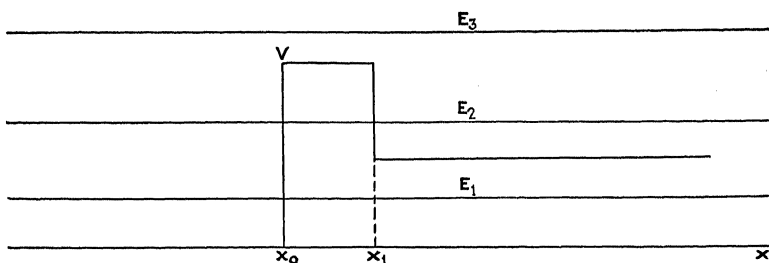


FIG. 63.—Potential barrier. The barrier is between x_0 and x_1 . Motion with the energy E_1 would have a wave function large only to the left of x_0 , rapidly decreasing to the right of x_0 . With the energy E_2 , the wave function would be large on both sides of the barrier, small but not zero within it, giving the possibility of penetrating the barrier. The wave function of energy E_3 would be large everywhere.

barrier. An electron of low energy, as E_1 , will be confined, except for a small exponential term, to the region to the left of x_0 , or the interior of the metal, and will never escape. An electron of the very high energy E_3 will be able to escape, either on classical or quantum mechanics. But an electron of intermediate energy E_2 can penetrate the barrier and escape on quantum mechanics, but not in Newtonian mechanics. These electrons of high energy, as E_2 or E_3 , are met only at high temperatures, so that we see the connection with thermionic emission.

Consider an electron of energy E_2 , and a solution which to the right of x_1 is a progressive wave traveling to the right. Then within the barrier we should have a combination of the two kinds of exponential functions, one increasing exponentially to the left, the other decreasing, with amplitudes properly chosen to satisfy the boundary condition of continuity of the function and its slope at x_1 . These in turn will join onto two progressive waves

to the left of x_0 , one traveling to the right, one to the left. The final result may be described as follows: An incident progressive wave falling on x_0 from the left; a reflected wave in the region to the left; a transmitted wave to the right of x_1 , the transmission through the barrier being of the real exponential form. We can tell something without much trouble about the amount transmitted. For within the barrier the term

$$\frac{\text{constant}}{\sqrt[4]{V-E}} e^{\frac{-2\pi}{h} \int \sqrt{2m(V-E)} dx},$$

increasing exponentially to the left, is the important one. And we readily see that its amplitudes at x_1 and x_0 measure, at least in order of magnitude, the relative amplitudes of transmitted and incident waves. Thus the fraction transmitted depends on the square of the quantity $e^{\frac{-2\pi}{h} \int_{x_0}^{x_1} \sqrt{2m(V-E)} dx}$. We work out examples of this integral in the problems, showing that there can be barriers of atomic size small enough so that appreciable penetration takes place, though in general this is not true, since a small increase in the height or breadth of a barrier can, on account of the exponential, make an enormous difference in the ease of penetration.

213. Motion in a Finite Region, and the Quantum Condition.—

Assume next that the kinetic energy is positive only in a finite region, so that classically the motion would be limited to that region. Then there will be a boundary condition on the wave function at each boundary of the region. Just as with the string held at both ends, this condition cannot in general be satisfied; it can be satisfied only for certain energies (corresponding to certain frequencies with the string). Using the approximate method of Wentzel, Kramers, and Brillouin, it is easy to see the nature of this condition. For each boundary must have essentially the treatment of Fig. 62, only the exponential decreasing toward infinity being allowed, whereas with an arbitrary energy the exponential would increase toward infinity in at least one direction. We have seen that the exponential part of the curve corresponds to $\frac{1}{8}$ wave length of the sinusoidal part. The num-

ber of wave lengths between x_1 and x_2 is $\int_{x_1}^{x_2} \frac{p}{h} dx$. Thus the whole number of waves between $-\infty$ and ∞ , taking account of the two exponential ends, is $\int_{x_1}^{x_2} \frac{p}{h} dx + \frac{1}{4}$. Since the function goes to zero

at both limits, this must be a whole number of half waves, or twice it must be a whole number. Hence

$$2 \int_{x_1}^{x_2} \frac{p}{h} dx + \frac{1}{2} = 1, 2, 3, \dots$$

$$2 \int_{x_1}^{x_2} p dx = \left(n + \frac{1}{2}\right)h, n = 0, 1, 2, \dots \quad (10)$$

This is the so-called quantum condition, developed particularly by Sommerfeld. We must remember that, since it is based on the approximation of Wentzel, Kramers, and Brillouin, it is not necessarily an exact condition. In some cases, as the linear oscillator, taken up in Prob. 5, it proves to be exactly true. In other cases, as a particle moving freely between two reflecting walls, as considered in Prob. 10, a similar condition holds, except that the quantum number, which here is $(n + \frac{1}{2})$, a half integer, is instead a whole integer. There are still other problems, as the hydrogen atom, in which a modified form of the condition is correct. In most cases, however, the quantum condition gives only an approximation, though a good one.

A number of problems can be solved exactly when the motion is confined to a finite region, and it is by comparison with these exact solutions that one can check the method of Wentzel, Kramers, and Brillouin, and the quantum conditions. Thus, in Prob. 5 we show that the wave equation for the linear oscillator can be solved as an exponential times a power series. This power series in general diverges for large x , indicating a function which goes to infinity as x becomes infinite. But if we give the energy particular values, which prove to be just those for which $2 \int_{x_1}^{x_2} p dx = (n + \frac{1}{2})h$, the series breaks off to form a polynomial, and the function goes to zero at infinity. These are the only solutions we can use, and they give just the quantum condition we found before, though by a quite different method. Again, a rotator, a solid of fixed moment of inertia and constant angular momentum rotating on an axis in the absence of torques, has a wave function $e^{\pm \frac{2\pi i}{h} \int p d\theta}$, where p, θ are angular momentum and angular rotation. Since p is constant, the real forms of this are \sin (or \cos) $(\pm 2\pi p \theta / h)$. For this to represent a single-valued function of position, it is necessary, as with the circular membrane, to have the function periodic with period 2π in θ . Thus we must

have $2\pi p/h = \text{integer} = m$, giving whole integral quantum numbers in this case, and determining the angular momentum as $m h/2\pi$.

214. Motion in Two or More Finite Regions.—In classical mechanics, we do not have to discuss specially the case where there are two separated regions where the kinetic energy is positive, separated and bounded by regions where it is negative; the motion occurs in one or the other of these regions, and that is the end of it. But in wave mechanics, the barrier between regions is not entirely impenetrable. We shall not go into the mathematical details of the solution, for, while they involve no new ideas, they are rather tedious. But the result is that the particles can penetrate the barrier and go from one region to the other, just as we have seen in a previous section in considering a barrier between two regions each extending to infinity in one direction. There are some new situations, however. Each region by itself would have stationary states of its own, if the other were not there. But with the two, no one of these states refers to motion wholly in the one region; the particle can go back and forth from one to the other. However, if the energy level is one that is characteristic of the one region and not of the other, the particle spends almost all of its time in that region of which its energy is characteristic. Once in a while it leaks over to the other side, but it soon finds its way back. It may be, however, that a given energy level will be characteristic of both regions; this is surely true if they are identical regions. Then the particle will travel back and forth from one to the other, spending equal lengths of time in each. This is an important physical case. For instance, in the hydrogen molecule, both atoms are just alike, and an electron finds a potential field which has two minima, one at each nucleus. It then can oscillate back and forth, spending half its time about one nucleus, half on the other. These problems are closely analogous to that of coupled oscillators, which we have already taken up. There we found that one oscillator would not move without setting the other into vibration, and similarly here the wave function cannot be large in one region without having a value in the other also. And here we have a special case if the two regions are identical, as we did before if the oscillators were equivalent. We shall find that the whole mathematical treatment is closely analogous.

We can finally have motion in two regions, one finite, the other reaching to infinity. Then, if the particle starts in the finite range, it is able in time to leak across the boundary, and go off to infinity. The present explanation of radioactivity is based on this idea. An alpha particle is supposed to be held in an atomic nucleus by a restoring force pulling it to a position of equilibrium. But if it were outside, then being positively charged, it would be repelled from the positive nucleus, the repulsion going to zero at infinity. Thus we should have a potential curve as in Fig. 64, where potential is drawn as function of r , the distance

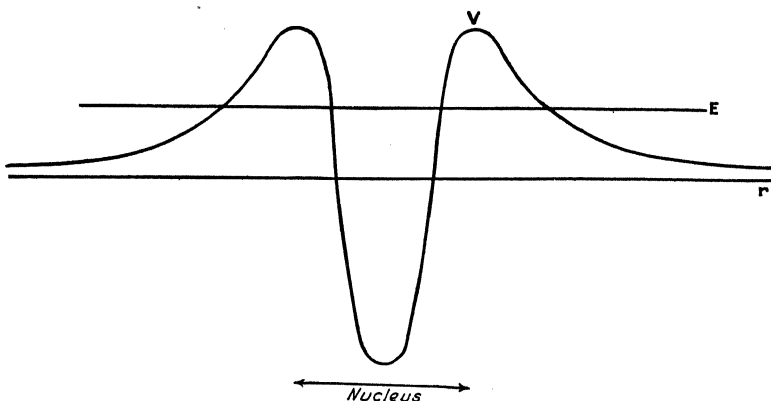


FIG. 64.—Potential curve for radioactive disintegration. A wave function of energy approximately E , starting out as a wave packet within the valley of the potential curve, would gradually leak out through the barriers.

from the center of the nucleus to the escaping alpha particle. If now the alpha particle has energy E , and is originally within the nucleus, it will eventually leak out, going off to infinity with a large kinetic energy, as the ejected alpha particles are actually found to have.

Problems

1. Prove that the function $\frac{\text{constant}}{\sqrt{E-V}} e^{\pm \frac{2\pi i}{h} \int p \, dx}$, where $p = \sqrt{2m(E-V)}$, is an approximate solution of Schrödinger's equation, becoming more and more accurate as V changes by a smaller and smaller amount in a wave length.
2. Note that in Bessel's equation for J_m , when $m > 0$, there is a region near the origin where the Wentzel-Kramers-Brillouin approximate solution is exponential rather than sinusoidal. Discuss the solution qualitatively for $x < m$, where m is fairly large, showing how this solution joins onto the sinusoidal one found in Prob. 9, Chap. XIV.

3. Note that the solution of Schrödinger's equation is sinusoidal or exponential in a region of constant potential. Discuss the one-dimensional problem of particles going from one region with constant potential V_1 to a second region of constant potential V_2 , when the energy is great enough so that the kinetic energy is positive in both regions. Satisfy boundary conditions at the surface, making u and its derivative continuous, joining the two sinusoidal functions together at the boundary. Show that some of the incident particles travel across the boundary, but that some are reflected back, contrary to classical mechanics. Find the fraction reflected.

4. Assuming the potential function of Fig. 63, consider particles striking the barrier from the left with energy E_2 . Set up the solution, satisfying the boundary conditions at x_0 and x_1 , and get an expression for the reflection coefficient as a function of the height of the barrier. Show that the reflection coefficient approaches unity if the barrier is infinitely high, or infinitely broad.

5. Show that the integral $2 \int_{x_1}^{x_2} p \, dx$ for an oscillator of natural frequency ν , energy E , equals E/ν . Show that therefore the quantum condition leads to the energy levels $E = (n + \frac{1}{2})h\nu$ for the oscillator.

6. Solve the problem of the linear oscillator of frequency ν , where $V = 2\pi^2\nu^2mx^2$. To do this, set

$$u = e^{-\frac{2\pi^2m\nu}{h}x^2} v(x),$$

and set up the differential equation for $v(x)$. For convenience, introduce the change of variables $y = 2\pi\sqrt{m\nu/h} \, x$. Solve in series, and show that the resulting series breaks off only if $E = (n + \frac{1}{2})h\nu$, where n is an integer.

7. Using the series of Prob. 6, investigate the behavior for large x if the series does not break off. Show that for very large x , v approaches the

series for $e^{\frac{4\pi^2m\nu}{h}x^2}$, so that the whole function u increases exponentially with x^2 , and cannot be used as a wave function.

8. Compute and plot wave functions of the linear oscillator corresponding to $n = 0, 1, 2, 3, 4$. From the graphs find the region in which the solution is oscillatory (that is, the region between the points of inflection). Draw the potential curve and the values of E corresponding to these four stationary states, and show that the motion is oscillatory in the region where the kinetic energy is positive.

9. Set up the approximate solution for the linear oscillator problem by the Wentzel-Kramers-Brillouin method, getting expressions for the functions in both the sinusoidal and the exponential ranges. Investigate to see how well these functions join on at the point of inflection.

10. Compute and plot the approximation of Prob. 9 corresponding to $n = 4$, and compare with the exact solution.

11. A particle executes one-dimensional motion in a container, having constant potential inside, and with the potential becoming suddenly infinite at the walls, so that the particle never gets out. Show that the boundary condition is that the wave function must be zero at the walls, as the displacement of a stretched string is zero at its ends. Find the wave functions of the problem, and find the energy of the particle in the n th stationary state.

CHAPTER XXX

THE CORRESPONDENCE PRINCIPLE AND STATISTICAL MECHANICS

The quantum condition has a close connection with the phase space and the Hamiltonian method, which we have discussed in Chap. IX. Hamiltonian methods have, in fact, been the guiding principle in the development of the quantum theory. At the same time, the phase space is fundamental to statistical mechanics, the mathematical foundation of thermodynamics. For that reason, we may profitably treat these subjects together, though, of course, statistical mechanics can be developed entirely from the basis of classical theory. Nevertheless, on account of the essentially statistical nature of the quantum theory, it yields an almost more natural approach to statistical mechanics than is possible in Newtonian mechanics, and by developing the two together we can illustrate the correspondence between classical and quantum mechanics which must hold, since the classical theory is a correct limiting form of quantum theory for large-scale problems.

215. The Quantum Condition in the Phase Space.—In Fig. 11, Chap. IX, we show the phase space for a linear oscillator, with a line of constant energy E , an ellipse of semiaxis $\sqrt{E/2\pi^2m\nu^2}$ along the q axis, and $\sqrt{2mE}$ along the p axis. These quantities measure the maximum coordinate and momentum, respectively, which a particle of E attains during its motion. For such an oscillator, the quantum condition (10), Chap. XXIX, equates twice the integral of $p dq$ between the minimum and maximum q values to $(n + \frac{1}{2})h$. Just as $\int y dx$ measures the area under the curve $y(x)$, so $\int p dq$ measures the area under the curve $p(q)$ in the phase space. The integral $\int_{q_1}^{q_2} p dq$ is that part of the area of the ellipse above the q axis, and to get the whole integral we double this, obtaining also the integral below the q axis. This may be written as an integral around the contour, from q_1 to q_2 around the upper branch of the curve, then back to q_1 along the

lower part of the curve, in which p and dq are both negative, so that we contribute an equal positive term to the integral. In other words, the quantum condition may be written

$$\oint p \, dq = (n + \frac{1}{2})h, \quad (1)$$

where \oint indicates an integral around the contour. And the physical interpretation is that the quantum integral is the area of the ellipse. Since this is πab , where a and b are the two semiaxes, it is $\pi\sqrt{2mE}\sqrt{E/2\pi^2m\nu^2} = E/\nu$, giving from Eq. (1) $E = (n + \frac{1}{2})h\nu$, in agreement with the result of Prob. 5, Chap. XXIX.

The results of the last paragraph are general: with any one-dimensional motion the quantum integral $\oint p \, dq$ represents the area of phase space enclosed by the path of the representative point, and the quantum condition says that this area is $(n + \frac{1}{2})h$, approximately. If we take successive stationary states, connected with successive quantum numbers n , each will have a curve in phase space, the path of a representative point of the corresponding energy, and the area between successive curves will, by the quantum condition, be h . Thus the phase space is divided up by these paths into a set of cells, each of area h , one for each stationary state.

216. Angle Variables and the Correspondence Principle.—We have seen in Chap. IX, Sec. 59, that a change of variables, called a contact transformation, can be set up, in which the new coordinate w increases uniformly with time, and the momentum J stays constant. To visualize this transformation in the case of the oscillator, we may imagine the phase space plotted with such scales of coordinates and momenta that the ellipses of constant energy become concentric circles. Then the new variables are essentially polar coordinates in phase space, the coordinate being the angle divided by 2π , the momentum being proportional to the square of the radius, so that obviously the angle variable increases uniformly with time, the momentum staying constant. The momentum J , called the action variable, proves in fact to be precisely the area of the ellipse, or circle, or the same integral $\oint p \, dq$ which we meet in the quantum condition.

In terms of the action variable, often called the phase integral, we saw that Hamilton's equation

$$\frac{\partial H}{\partial J} = \frac{dw}{dt} = \nu \quad (2)$$

gave the frequency in terms of a simple calculation. This formula permits us to make an extremely interesting connection between the classical frequency of motion of a system and the frequency of the light emitted in a transition between two states of energy E_2 and E_1 according to Bohr's frequency condition

$$E_2 - E_1 = h\nu, \quad (3)$$

described in Sec. 201. On the quantum theory, most energies H of the system are not allowed; we may have rather only those satisfying the quantum condition (1). Thus H cannot be regarded as a continuous function of J . We may, however, replace the derivative $\partial H/\partial J$ of (2) by the difference ratio $\Delta H/\Delta J$, in which ΔH is the energy difference between two states, ΔJ the difference between their phase integrals. If we choose two states whose quantum numbers differ by unity, we have $\Delta J = h$, so that the difference ratio is

$$\frac{(E_2 - E_1)}{h} = \nu,$$

giving precisely the quantum frequency according to Eq. (3). Hence we have the following relation: the derivative $\partial H/\partial J$ gives the classical frequency of motion of a system; the difference ratio $\Delta H/\Delta J$, where the difference of J is one unit, gives the frequency of emitted light according to the quantum theory, or the frequency of oscillation of the oscillator mentioned in Sec. 202. We shall consider later the significance of transitions of more than one unit in J .

For the oscillator, as one can immediately see from the fact that its energy in the n th state is $(n + \frac{1}{2})h\nu$, the classical and quantum frequencies are exactly equal, the derivative equaling the difference. This is plain from the fact that here $E = J\nu$, so that the curve of E against J is a straight line, and the ratio of a finite increment in ordinate, divided by a finite increment in abscissa, equals the slope or derivative. But for any other case the curve of E against J is really curved, so that the derivative and difference ratio are different, and classical and quantum frequencies do not agree. Thus in Fig. 65 we show an energy curve for an anharmonic oscillator, in which the tightness of binding decreases with increasing amplitude, the frequency decreases, and therefore the slope decreases with large quantum numbers. Here the classical frequency, as given by the slope of the curve, does not agree with the quantum frequency con-

ned with the transition indicated, from $\frac{5}{2}h$ to $\frac{3}{2}h$, for the quantum frequency is the slope of the straight line connecting E_2 and E_1 . We may assume, however, that if we go to a very high quantum number, so that we are far out on the axis of abscissas, any ordinary energy curve will become asymptotically fairly smooth and straight, so that the chord and tangent to the curve will more and more nearly coincide. This certainly happens in the important physical applications we shall make. In these cases, we may state Bohr's correspondence principle:

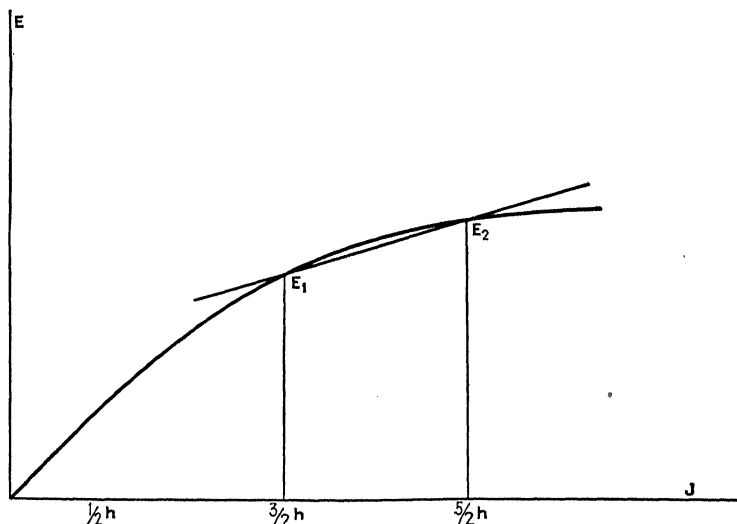


FIG. 65.—Energy curve for anharmonic oscillator. Slope of curve gives classical frequency, slope of straight line connecting E_2 and E_1 gives quantum frequency.

in the limit of high quantum numbers, the classical and quantum frequencies become equal. This is essentially simply a special case of the general result stated in Chap. XXVIII, that in the limit of small wave lengths (which for most practical purposes is the same as the limit of high quantum numbers) the classical and quantum theories become essentially equivalent.

217. The Quantum Condition for Several Degrees of Freedom.

In classical mechanics, we have seen that certain problems, like the two-dimensional oscillator and the central field motion, are separable, so that they can be broken up into several one-dimensional motions. Since each of these motions was periodic, the

whole motion is multiply periodic in these cases. In these particular problems with several degrees of freedom, separation of variables can also be carried out in the quantum theory. In phase space we can pick out the two-dimensional space representing one coordinate and its conjugate momentum, and the projection of the representative point on this plane will trace out a closed curve. There is a quantum condition associated with this coordinate, the area enclosed by the curved path in the two-dimensional space being a half integer times h . Thus we have a quantum number associated with each degree of freedom in such a problem. Further, we can introduce angle and action variables connected with each of the coordinates, just as if each formed a problem of one degree of freedom. The various frequencies of the multiple periodicity can be found by differentiating the energy with respect to the various J 's, and the correspondence principle can be applied to connect these classical frequencies with the quantum frequencies associated with various possible transitions.

It can be shown in general that any coordinate of, say, a doubly periodic motion, can be analyzed into a sort of generalized Fourier series in the time, in which terms appear of frequencies

$$\tau_1\nu_1 + \tau_2\nu_2, \quad (4)$$

where τ_1, τ_2 are arbitrary integers. This is the generalization of the ordinary Fourier representation for a purely periodic motion, in which all frequencies $\tau_1\nu_1$ will in general appear which are integral multiples of the fundamental frequency. Now we can carry out a general correspondence between any one of these overtone or combination frequencies and a corresponding transition. Thus let us consider the transition in which J_1 changes by τ_1 units, J_2 by τ_2 units, where J_1 and J_2 are the two action variables. The quantum frequency emitted will be

$$\frac{E(J_1, J_2) - E(J_1 - \tau_1 h, J_2 - \tau_2 h)}{h}, \quad (5)$$

where E is the energy, written as function of the J 's. But if we are allowed to replace differences by derivatives, as we assume we are in the correspondence principle, this becomes

$$\frac{1}{h} \left(\frac{\partial E}{\partial J_1} \tau_1 h + \frac{\partial E}{\partial J_2} \tau_2 h \right) = \tau_1 \nu_1 + \tau_2 \nu_2,$$

in agreement with Eq. (4), if $\nu_1 = \partial E / \partial J_1$, $\nu_2 = \partial E / \partial J_2$. Thus we have a one to one correspondence between all possible overtone vibrations of the classical motion and all possible quantum transitions. This correspondence is of great importance, for instance, in discussing intensities of radiation, as we shall see later. For each component of the Fourier representation is a sinusoidal vibration, with frequency (4), and a certain amplitude A_{τ_1, τ_2} . This oscillation, if it were the oscillation of an electric charge, would send out a radiation of frequency (4), with an intensity proportional to the square of the amplitude, as we have seen in Chap. XXV, where we found a rate of radiation $\frac{16\pi^4 A^2 \nu^4}{3c^3}$. Thus this Fourier component A would directly

determine the intensity of classical radiation. It then seems very reasonable that, at least in the limit of high quantum numbers, this intensity would agree approximately with the intensity of the corresponding quantum transition given by Eq. (5). Thus one can derive from correspondence principle definite information about probabilities of quantum transitions, for the rate of radiation of energy in a particular transition is proportional to the number of transitions occurring per unit time, or the probability of transition. We shall return to this question in a later chapter.

The results which we have mentioned are all for multiply periodic, separable problems in several dimensions. With an n -dimensional problem, and a $2n$ -dimensional phase space, there are n J 's which stay constant during the motion. Thus we may set up n sets of surfaces, $J_1 = \text{constant}$, $J_2 = \text{constant}$, . . . $J_n = \text{constant}$, in the phase space, and the representative point moves so that it stays on an intersection of all n surfaces, or in an n -dimensional region, instead of all through the $(2n - 1)$ dimensional energy surface, as it would in quasi-ergodic motion. The particular surfaces $J_1 = (n_1 + \frac{1}{2})h$, $J_2 = (n_2 + \frac{1}{2})h$, etc., divide up the phase space into cells, each of which is seen to have the volume h^n , at least in simple cases, and a little examination shows that there is just one stationary state per cell. Of course, the path of a representative point is always on an energy surface, and if we take only the quantized J values, the corresponding representative points lie only on the energy surfaces corresponding to quantized energy values. In many cases it proves to be true that a number of different stationary states have the same energy. Such a problem is called degenerate, and the number of

different states connected with the energy level is called the a priori probability of the level. In such a case the volume of phase space between this energy surface and the next adjacent one proves to be h^n times the a priori probability.

For a quasi-ergodic system, as we have said, there are no quantities like the J 's which stay constant, other than the energy. There are still stationary states in the quantized problem, though they are not determined by ordinary quantum conditions. They are derived from solutions of the Schrödinger equation, however, and the boundary conditions lead to definite stationary states, as with one-dimensional motion. Thus we can always introduce energy surfaces in the phase space, corresponding to the quantized states. Generally quasi-ergodic systems are not degenerate, all energy levels being distinct, and the volume of phase space between successive energy levels will always be, at least to an approximation, equal to h^n . These relations prove to be of importance in investigating the statistical mechanics of collections of systems in the phase space.

218. Classical Statistical Mechanics in the Phase Space.—In Chap. IX we have investigated the motion of a representative point in the phase space. Statistical mechanics, however, like any statistical science, deals not with single points but with an enormous number of individuals, investigating their average behavior. In its applications to thermodynamic problems, there are two principal methods, both of which are frequently used. In the first of these, we deal, for instance, with a gas composed of a great many identical molecules. These molecules themselves form the individuals whose average properties we investigate. Thus the phase space we use is one in which there are enough coordinates and momenta to describe a single molecule. Such a space is often called a μ space. The second method is more powerful but more abstruse: the individuals with which we deal are whole systems, as whole samples of gas, and we imagine a large collection, often called an ensemble, or assembly, of such samples, all just alike in such gross properties as volume, temperature, and density, but differing in their finer details, as the positions and velocities of individual atoms or molecules. These might represent different pictures of the same gas at different times; or they might represent different repetitions of the same experiment, all controllable conditions being held fixed. Finding averages over such ensembles means then finding the time aver-

age, or finding the average obtained by repeating the experiment many times. The phase space required for this second method has as many coordinates and momenta as there are in the whole system, a very great number if the system contains many molecules. This space is often called the Γ space. As to the distinction between the methods of the μ and the Γ spaces, the general situation is that they are equivalent when applied to perfect gases; but if the molecules interact, they can no longer be treated as independent systems and described by separate points in the μ space, but one must instead consider the whole system together, and use the Γ space. The latter method is then the one which we shall use more often. Both methods are alike, however, in using phase spaces, and in considering the motion of a swarm of points in such a space.

We imagine an ensemble of a great many, or even an infinite number, of points in a phase space. As time goes on, with the points moving, the effect is as of the whole swarm flowing, like a liquid or gas composed of atoms. In fact, many of the ideas of hydrodynamics can be applied in this case, as we shall show in the next section. We introduce first the density of points as a function of the p 's and q 's:

$$f(p_1 \dots p_n, q_1 \dots q_n) dp_1 \dots dp_n dq_1 \dots dq_n$$

gives the number of points in the $2n$ -dimensional volume element $dp_1 \dots dp_n dq_1 \dots dq_n$. The velocity of points in the phase space is then given by Hamilton's equations, $dq_i/dt = \partial H/\partial p_i$, $dp_i/dt = -\partial H/\partial q_i$, as we pointed out in Sec. 52, Chap. IX. Thus we have the necessary quantities to describe the motion of the points as a flow, and in the next section we apply the equation of continuity and investigate its consequences.

219. Liouville's Theorem.—Consider the steady flow of a fluid of density ρ , velocity v . The equation of continuity is $\partial \rho/\partial t + \text{div}(\rho v) = 0$, or $\partial \rho/\partial t + \rho \text{div } v + v \cdot \text{grad } \rho = 0$, (6) if the density varies from point to point. We are interested particularly in a divergenceless flow, for which $\text{div } v = 0$, for it turns out that the flow of points in the phase space is of this sort. It is easy to see that this corresponds to the flow of an incompressible fluid. For let us find $d\rho/dt$, the time rate of change of density with time. This is given by

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \frac{\partial \rho}{\partial x} \frac{dx}{dt} + \frac{\partial \rho}{\partial y} \frac{dy}{dt} + \dots = \frac{\partial \rho}{\partial t} + v \cdot \text{grad } \rho, \quad (7)$$

where $d\rho/dt$ is the rate at which density changes if we follow along with a particle of fluid. But now if $\text{div } v = 0$, Eq. (6) becomes $d\rho/dt = 0$, showing that the density following the particle does not change with time, which is to be expected if the fluid is incompressible. This does not imply, however, that the density of the fluid is at all points the same. Let us imagine a fluid composed of large droplets of one sort of fluid suspended in another. If the fluids are chosen so that they do not mix, and the surfaces of separation remain sharp, then the density will change from point to point, as we go from the one fluid to the other. Further, if the whole fluid is moving, the density at any point of space will change with time, as first the one sort of fluid, then the other, will be carried past this point. But if the fluid is incompressible, the density of a particular part of the fluid, as we follow it in its motion, will be constant. That is, $v \cdot \text{grad } \rho$ and $\partial\rho/\partial t$ are separately different from zero, but their sum vanishes.

The situation we have just described holds for the motion of points in the phase space. The $2n$ -dimensional velocity of points, as we have seen in the last section, has components dq_i/dt , dp_i/dt , where i goes from 1 to n . Then the analogue to the divergence is

$$\begin{aligned} \text{div } v &= \frac{\partial}{\partial q_1} \frac{dq_1}{dt} + \frac{\partial}{\partial q_2} \frac{dq_2}{dt} + \cdots + \frac{\partial}{\partial p_1} \frac{dp_1}{dt} + \cdots \\ &= \frac{\partial}{\partial q_1} \frac{\partial H}{\partial p_1} + \frac{\partial}{\partial q_2} \frac{\partial H}{\partial p_2} + \cdots - \frac{\partial}{\partial p_1} \frac{\partial H}{\partial q_1} - \cdots = 0. \quad (8) \end{aligned}$$

Thus on account of Hamilton's equations the flow is divergenceless. Then we see that the flow is an incompressible flow, the density of points remaining constant as we follow a particle. This is Liouville's theorem.

220. Distributions Independent of Time.—The principal use of distributions in the phase space is for thermodynamic purposes, and here we are interested in thermal equilibrium, and in distributions independent of time. An ensemble independent of time is one for which $\partial\rho/\partial t = 0$. To get that, we see from Eq. (7) that we must have $v \cdot \text{grad } \rho = 0$. This means that the rate of change of density along the direction of flow, or along the streamline, is zero. In other words, all along a single line of flow, or through a single tube of flow of infinitesimal cross section, the density is constant. We may imagine the whole phase space divided up into tubes of flow, and then any distribu-

tion in which each tube has its own constant density through its whole volume, no matter how this density may change from one tube to another, will be independent of time.

In a multiply periodic motion, the lines of flow will be given by $J_1 = \text{constant}$, $J_2 = \text{constant}$, \dots . Thus if we make the density any arbitrary function of the J 's, we shall have a distribution independent of time. Remembering that the density is the function f , this is

$$f(p_1 \dots p_n, q_1 \dots q_n) = F(J_1 J_2 \dots J_n). \quad (9)$$

On the other hand, in a quasi-ergodic motion, a single line of flow will in time come arbitrarily near to every point of an energy surface. Thus the only distribution which will be independent of time in this case is one in which the density is constant all over an energy surface:

$$f(p_1 \dots p_n, q_1 \dots q_n) = F(E). \quad (10)$$

Of course, the ensemble (10) would be independent of time even in a multiply periodic motion, but it is more specialized than is necessary in that case. For instance, in an ensemble of systems each consisting of a particle in central motion, we could make an ensemble in which all parts of the phase space corresponding to the same energy had the same density, and this would be constant. But we could equally well make the density in different parts of the space corresponding to the same energy but different angular momentum different, and still, as long as the angular momentum was conserved, this distribution would be constant. Any perturbation which involved slow changes of angular momentum, however, would destroy the constancy of this distribution, whereas if we had started with one which depended only on energy, it would not be affected by such a perturbation.

The ordinary systems which we deal with thermodynamically are assumed to be so complicated that they are quasi-ergodic. Thus the only type of ensemble independent of time is that of Eq. (10), in which the density is a function of the energy. This is the sort which we shall consider in thermodynamic applications.

221. The Microcanonical Ensemble.—A particularly important ensemble is that called the microcanonical ensemble, in which all the systems of the ensemble have practically the same energy. More precisely, we have

$$\begin{aligned}
 f(p_1 \cdots p_n, q_1 \cdots q_n) &= F(E) \\
 &= \text{constant for } E_0 < E < E_0 + \Delta E \\
 &= 0 \text{ otherwise.}
 \end{aligned} \tag{11}$$

It is evident that an arbitrary ensemble can be made up by superposing microcanonical ensembles, the ensemble whose systems lie between E_0 and $E_0 + \Delta E$ having a constant density so chosen as to give the proper density in that particular part of the energy space. In thermodynamics the microcanonical ensemble is often used, when we wish to deal with the statistical properties of systems at a given temperature, for energy content is correlated with temperature in such a way that systems of the same temperature have just about the same energy, and therefore are represented at least approximately by a microcanonical ensemble.

222. The Canonical Ensemble.—More suitable than the microcanonical ensemble for discussing temperature equilibrium proves to be a slightly different one called the canonical ensemble. In this distribution the density function is given by

$$f = F(E) = \text{constant } e^{-\frac{E}{kT}}, \tag{12}$$

where E is the energy, k a constant, called Boltzmann's constant, equal to 1.37×10^{-16} c.g.s. units, T is the absolute temperature. We shall discuss in a later chapter the particular properties of this ensemble, and its advantages. This ensemble has not only the property of remaining unchanged with time, if the system is left to itself, but also of remaining unchanged if the system can interchange energy with another of the same temperature. This is evidently necessary for thermal equilibrium, and the canonical ensemble is the only one in general which has this property. From this ensemble we can derive interesting results, though we mention only a few. We may, for instance, use the μ space, each system being a molecule. The energy of such a molecule is $\frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) + V$, so that the probability of finding a molecule having its coordinates and momenta within the limits x_0 and $x_0 + dx$, y_0 and $y_0 + dy$, z_0 and $z_0 + dz$, p_{x0} and $p_{x0} + dp_x$, etc., is proportional to

$$e^{-\frac{\frac{1}{2m}(p_{x0}^2 + p_{y0}^2 + p_{z0}^2) + V}{kT}} dx dy dz dp_x dp_y dp_z. \tag{13}$$

This law is ordinarily called the Maxwell-Boltzmann distribution law. From it we can easily find that the velocities are distributed according to Maxwell's distribution of velocities, and that the density in ordinary space at different points is proportional to $e^{-v/kT}$. We leave these proofs for problems. If on the other hand we use the Γ space, E represents the energy of the whole sample of gas, and we can prove easily that the energy of an individual sample in the ensemble is very nearly the same as that of any other sample. Thus for such a system the canonical ensemble is very similar to the microcanonical ensemble. One gets the same thermodynamic results using either ensemble, but the canonical ensemble is both more correct theoretically and decidedly simpler for most of its applications.

223. The Quantum Theory and the Phase Space.—In Sec. 210, Chap. XXIX, we have seen that a stationary state of a one-electron problem corresponds to a classical particle whose energy is determined, but whose initial time of starting is undetermined. More accurately, it corresponds to an ensemble of particles, all of the same energy, but with phases distributed in such a way that the properties of the ensemble are independent of time. This, however, is exactly a microcanonical ensemble. This may be connected with the uncertainty principle for energy, Eq. (2) of Chap. XXVIII, which states that the uncertainty of energy multiplied by the uncertainty of time is equal to h . If then we set up an ensemble of particles all of exactly the same energy, it must necessarily be true that the uncertainty of time of one of the particles is infinite. That is, we know nothing at all as to its phase, or the ensemble consists of particles in all possible phases. And since it is a stationary state we are dealing with, nothing depends on time. In other words, with the quantum theory, the mere process of setting up a stationary state automatically sets up a microcanonical ensemble. We need not do that specially, and we need not prove Liouville's theorem to find out how to get ensembles independent of time. In this way the quantum mechanics is more convenient for statistical purposes than classical mechanics.

With problems with several variables, the stationary state certainly represents an ensemble independent of time. If the problem is multiply periodic, it will represent an ensemble of states all of the same J values (that is, the same set of quantum numbers), but arbitrary phases. On the other hand, if it is

quasi-ergodic, it will represent a microcanonical ensemble. And even in a multiply periodic, degenerate case, where there are several stationary states of the same energy, we can always set up a microcanonical ensemble, by combining all the various states of the same energy. Each one of these states will correspond to a volume h^n of the phase space. Then if the microcanonical ensemble is to have a constant density of points over a region between two energy surfaces, it will have a definite number of points for each element of volume h^n , and hence a constant and equal number of points for each of these substates of the same energy. We may say that in this ensemble the number of systems in any group of substates is proportional to the a priori probability of this group of states; that is, simply proportional to the number of substates in the group.

The distribution function $f(p_1 \cdot p_n, q_1 \cdot q_n)$ for the quantum theory involves us in rather complicated considerations, which we shall take up in the next chapter. The reason is that the probability function which we are given directly is the square of the wave function, $\Psi\Psi$, and that is a function of the coordinates only, giving the probability of finding the coordinates within certain limits, independent of the momenta. In Sec. 210, Chap. XXIX, we have shown that this probability function approximately agrees with that found in classical mechanics. We postpone other comparisons between the quantum and classical distributions. But there is one feature of the quantum distribution function which should be mentioned at the outset. We have spoken above as if one could draw the paths of particles, and set up distribution functions, in the phase space, for the quantum theory as for the classical theory. But this is really not possible, as we can see from the uncertainty principle. This says that the uncertainty in the coordinate of a particle, multiplied by the uncertainty of its momentum, is of the order of magnitude of h . This product of uncertainties is simply an area in phase space. Instead of representing the particle by a sharp point, we can visualize it as a region in phase space, of dimensions Δq and Δp along the two axes. By the uncertainty principle, the area of this region is h . If we had the same thing in a number of dimensions, as n variables, the $2n$ -dimensional volume associated with the uncertain position and momentum of the particle or representative point would be h^n , just the volume associated with a stationary state. As a result of this

uncertainty, we must always be cautious about using the ideas of definite paths of representative points in the phase space. It would perhaps be more accurate to think of the paths, and energy surfaces, as having definite thicknesses, as if the point carried along its volume h^n , and allowed that to trace out a finite region of phase space.

The canonical ensemble can be set up in quantum theory as in classical mechanics. In the classical theory, it is the ensemble in which the number of points per unit volume is proportional to $e^{-E/kT}$. In quantum theory, the number of points in volume h^n , or the number in a given stationary state, is proportional to $e^{-E/kT}$, or this exponential is proportional to the probability of finding a system, chosen at random from the ensemble, in the stationary state in question. If we group together a number of degenerate substates all of energy E , and if there are g of them, so that the a priori probability of the group is g , the number of systems in the group is proportional to $ge^{-E/kT}$.

Problems

1. Take the problem of a particle executing one-dimensional motion in a container with constant potential inside, but impenetrable walls, as in Prob. 11, Chap. XXIX. Plot the path of the representative point in phase space, find the phase integral, and show that the quantum condition leads to the same stationary states and energy levels that were determined previously.

2. For the system of Prob. 1, compare (a) the frequency of oscillation of the particle back and forth between the walls, as determined classically by elementary argument; (b) the same frequency as determined by the formula $\partial H/\partial J$; (c) the emitted frequency on the quantum theory.

3. Draw the phase space for a rotator, as described in Sec. 213, and verify the quantum condition stated there.

4. Apply the correspondence principle to the radiation from a linear oscillator. Show that the Fourier components of the classical motion are zero corresponding to all transitions except those in which the quantum number changes by one unit only. From this one may infer that in the quantum theory only this particular transition can occur, the probability of any other sort of transition being zero. Such a result is called a selection principle.

5. Consider the motion of Prob. 6, Chap. IX, in which a particle executed simple harmonic motion on a rotating turntable. Assume that one quantum number, and phase integral, is associated with the rapid frequency of oscillation, and the other phase integral with the slower frequency of rotation of the turntable. From the Fourier analysis of the x component of motion, show that the only allowed transitions are those in which each quantum number changes by ± 1 unit. Show further that both must change together, there being no transitions of one quantum number alone, but that a transi-

tion of $+1$ unit in one of the quantum numbers is equally likely to be connection either with $+1$ or -1 of the other.

6. Find Maxwell's distribution of velocities, stating that the number of molecules of a gas for which the velocity is between v and $v + dv$ is proportional to

$$v^2 e^{-\frac{mv^2}{2kT}} dv.$$

To do this, use μ space, assume the Maxwell-Boltzmann distribution law. Consider a fixed point of space, so that x, y, z are constant, and we need only consider the three-dimensional momentum space. Note that the velocity is proportional to the radius $\sqrt{p_x^2 + p_y^2 + p_z^2}$ in the momentum space. The number of molecules between v and $v + dv$ is then proportional to the density of molecules in the momentum space, which from the Maxwell-Boltzmann law is constant for constant v , times the volume of momentum space between v and $v + dv$, which can be computed from the ordinary geometrical relations of a sphere. Determine the constant factor in the law so that your formula will give directly the fraction of all molecules in the range dv .

7. Find the mean kinetic energy of a molecule at temperature T . Note that the mean of any quantity $F(p, q)$ is given by

$$\bar{F} = \frac{\int F(p, q) f(p, q) dp \cdots dq \cdots}{\int f(p, q) dp \cdots dq \cdots},$$

where $f(p \cdots q \cdots)$ is the density function in the phase space, and the integration is over all parts of the phase space. Note also that since in this case F depends only on the momentum, the integrals in numerator and denominator can be factored into one integral over the momenta, one over the coordinates, and that the latter cancel out.

8. By integrating over all momenta, show that the space density of molecules in a gas is proportional to $e^{-V/kT}$. Apply this to the density of the atmosphere in the earth's gravitational field, assuming constant temperature. Find from this the rate of decrease of barometric pressure with altitude, at the earth's surface, assuming a reasonable atmospheric temperature.

9. In the Γ space, consider a canonical ensemble of N identical molecules, where N is very large. Assume that no forces act. Find the number of systems of the ensemble for which the total energy is between definite limits E and $E + dE$. To do this, note that the energy is proportional to $p_{x1}^2 + p_{y1}^2 + \cdots + p_{zN}^2$, or the square of the N -dimensional radius of the momentum space, so that the part of the space between E and $E + dE$ is the region between two corresponding hyperspheres. Note that the "volume" of a two-dimensional "sphere" (a circle) is πr^2 ; of a three-dimensional one, $\frac{4}{3}\pi r^3$; of an N -dimensional one, constant times r^N . Also note that the volume between r and $r + dr$ is $\frac{d(\text{volume})}{dr} dr$.

10. Show that the fraction of all systems is a canonical ensemble for which energy is between E and $E + dE$ is approximately given by a Gaussian error curve, $Ae^{-c(E-a)^2}$. Find c and a . (Hint: The function found in

Prob. 9 has a very sharp maximum, to be approximated by the error curve above. Expand the logarithm of the function in power series about its maximum, a , so that the logarithm equals constant $-c(E - a)^2 + \dots$, where there is no first power term because the expansion is about the maximum, and higher power terms than the second are to be neglected. Then the function is the logarithm of this power series, giving the value above. Show that the third and higher power terms are negligible unless $E - a$ is so large that the function itself has sunk to a negligible value.)

11. In the distribution of Prob. 10, show that the mean energy of the systems of the ensemble is just N times the energy of a single molecule, as found in Prob. 7. To get an idea of the range of distribution of energy about this mean, find the energy for which the Gaussian distribution curve falls to half its maximum value. Show that the energy difference between this value and the mean increases proportionally to \sqrt{N} , but that the percentage deviation of the energy, or the deviation divided by the total energy, goes down as $1/\sqrt{N}$, so that for large N the percentage deviation is extremely small.

CHAPTER XXXI

MATRICES

Suppose we have a problem, like the linear oscillator, in which there are no motions which go to infinity; that is, in which every motion is quantized, so that only discrete energy values are allowed. Let the n th energy value be E_n , the corresponding wave function u_n . Then a general solution of the wave problem, involving the time, is

$$\psi = \sum_n c_n e^{-\frac{2\pi i E_n t}{h}} u_n(x, y, z), \quad (1)$$

where we choose the negative exponential for reasons which will appear later. This function will shortly be derived as the solution of a wave equation involving the time, though we have not yet written down that equation. Now let us recall the meaning of ψ . It is the amplitude of a wave whose intensity gives the probability that the particle be found at a given place at a given time. Since ψ is complex, this intensity is given by multiplying by its conjugate; hence $\bar{\psi}\psi$ gives the desired probability. Or more precisely, the probability that the particle, at time t , is in the volume element $dx dy dz$, is $\bar{\psi}\psi dx dy dz$. One result appears at once from this: the probability that the particle be somewhere is unity, and this must be the sum of the probabilities that it be in all separate parts of space, or the integral of the probability over all space:

$$\iiint \bar{\psi}\psi \, dx \, dy \, dz = 1. \quad (2)$$

Now having the probability, we can proceed to get statistical information about the behavior of the particle.

224. Mean Value of a Function of Coordinates.—As we have seen in the last chapter, the first step in a statistical investigation is to find a distribution function. There we were interested in functions of coordinates and momenta of a particle or system, and we had a function $f(q_1, \dots, q_n, p_1, \dots, p_n)$, such that $f dq_1 \dots dp_n$ gave the number of systems having coordinates

and momenta in the range $dq_1 \dots dp_n$. To find the average of any quantity, given such a distribution, we proceed as follows: if the quantity is $F(q_1 \dots p_n)$, a function of coordinates and momenta, we multiply the function by the fraction of systems having those particular q 's and p 's, and integrate over all q 's and p 's. This fraction is $\frac{\int \bar{\psi} \psi dxdydz}{\int \int dq_1 \dots dp_n}$, so that the result is

$$\bar{F} = \frac{\int F \int \bar{\psi} \psi dxdydz}{\int \int dq_1 \dots dp_n} \quad (3)$$

where we denote the average of F by \bar{F} , to avoid confusion with the single bar indicating complex conjugates. Similarly in the present case we have a function $\bar{\psi}\psi$ which is a distribution function as far as coordinates are concerned: $\bar{\psi}\psi dxdydz$ is the probability (directly, since $\int \bar{\psi} \psi dxdydz = 1$) that the particle have coordinates within $dxdydz$. Thus if we have a function $F(x, y, z)$ of the coordinates, and wish its mean value, we have

$$\bar{F} = \int F \bar{\psi} \psi dxdydz = \int \bar{\psi} F \psi dxdydz, \quad (4)$$

where we prefer the latter method of writing it because it fits in with formulas which we shall have later. This does not tell us how to find averages of functions of the momenta, such as for example the energy; that is more complicated, and will be discussed in a later section. But we may wish, for instance with an atom or molecule, to find the mean value of the center of gravity, or moment of inertia, or some such function of position alone, and the formula suffices to determine it.

It is now very interesting to substitute our expansion of ψ in the expression for a mean value. That gives

$$\begin{aligned} \bar{F} &= \sum_{n,m} \bar{c}_n c_m e^{\frac{2\pi i}{h}(E_n - E_m)t} \int \bar{u}_n F u_m dxdydz \\ &= \sum_{n,m} \bar{c}_n c_m e^{\frac{2\pi i}{h}(E_n - E_m)t} F_{nm}, \end{aligned} \quad (5)$$

where by definition $F_{nm} = \int \bar{u}_n F u_m dxdydz$. The quantities F_{nm} form a two-dimensional array of numbers, of the sort known in mathematics as a matrix, and the individual F_{nm} 's are called matrix components.

225. Physical Meaning of Matrix Components.—Suppose we have an electron in an atom, and try to find its electric moment as a function of time; that is, its charge e times the displace-

ment of the electron, x . In other words, we wish the mean value of ex , which is

$$\overline{ex} = \sum_{n,m} \bar{c}_n c_m e^{\frac{2\pi i}{h}(E_n - E_m)t} (ex)_{nm}. \quad (6)$$

We observe that in the mean moment the terms depend on time through the expression $e^{\frac{2\pi i}{h}(E_n - E_m)t}$, having the frequency $(E_n - E_m)/h$. But this is just the frequency which by the quantum theory the atom should emit in jumping between the energy levels m and n . Hence we connect this particular matrix component with this transition. By the correspondence principle, in Sec. 217, Chap. XXX, we have already seen that associated with each transition there is a classical frequency of oscillation, and a corresponding Fourier component of the motion. It can now be shown that this Fourier component, in the limit of large quantum numbers, becomes equal to the matrix component $(ex)_{nm}$ of the electric moment, which appears in Eq. (6). The individual terms of Eq. (6) act like oscillators, radiating energy, and it proves to be true, though it requires a difficult analysis to show it, that the rate of radiation of the oscillator determines exactly the quantum probability of transition. For example, if a matrix component is zero, there will be no radiation of the corresponding frequency, no transitions are possible between the stationary states concerned, and we have what is called a selection principle, a principle selecting out certain transitions which can occur, the rest being forbidden.

The matrix components which we have noticed have been those where m and n were different. If we make a scheme of matrix components like

$$\begin{array}{cccc} F_{11} & F_{12} & F_{13} & \dots \\ F_{21} & F_{22} & F_{23} & \dots \\ F_{31} & \dots & & \\ \dots & & & \end{array} \quad (7)$$

we see that the components F_{11} , F_{22} , etc., along the principal diagonal all have $m = n$, so that our components with $m \neq n$ are just the nondiagonal components. The diagonal components, however, have a different meaning. They refer to time average properties of the system, rather than to the sinusoidal properties which are connected with radiation. Thus if we take the time average of \bar{F} (where the averaging in \bar{F} refers to

averaging over the probability distribution, not over time), the exponential term $e^{\frac{2\pi i}{h}(E_n - E_m)t}$ averages to zero, unless $n = m$, in which case it is unity. Hence we have

$$\text{time average of } \bar{F} = \sum_n \bar{c}_n c_n F_{nn}, \quad (8)$$

the double sum reducing to a single sum. Here, as we said above, only the diagonal components of the matrix of F appear.

We can understand this formula better if we notice the meaning of the c 's. To get at this, we observe that the c 's are the amplitudes by which the various overtones are multiplied, in order to get the whole wave function. Thus the quantities $\bar{c}_n c_n$, the squares of these amplitudes (taking account of the fact that they may be complex by multiplying by the conjugate) are quantities proportional to the intensities of the various overtones; and the interpretation of this is that they are proportional to the probability that the particle be in a given stationary state. As a matter of fact, we shall soon show that $\bar{c}_n c_n$ represents just the probability itself, the sum of all the probabilities, $\sum_n \bar{c}_n c_n$, being unity. Thus the formula

$$\text{time average of } \bar{F} = \sum_n \bar{c}_n c_n F_{nn}$$

means that F_{nn} is the time average of F over the n th stationary state, and $\bar{c}_n c_n$ the probability of finding the system in this stationary state, so that we multiply together and add to get the average over all stationary states.

226. Initial Conditions, and Determination of c 's.—Just as with the problem of the vibrating string, we may have initial conditions: we may know that the distribution ψ has a certain value at $t = 0$. Let us take a specific example: we may know that at $t = 0$ the particle is inside a given small volume, though we do not know where in that volume. Then we may ask as to its probable later motion. That is, we know that $\psi(x, y, z, t)$ is zero, at $t = 0$, outside the small volume, and has a constant value, or at least a sinusoidal form with constant amplitude, inside the volume. Now at $t = 0$, the exponentials become unity, so that we have $\psi(x, y, z, 0) = \sum_n c_n u_n(x, y, z)$. But this is just the familiar problem of expanding an arbitrary function

of x, y, z in a series of functions u_n . These are orthogonal functions; they are solutions of Schrödinger's equation, which is of the type already discussed in Prob. 10, Chap. XV, where we showed in general that the solutions were orthogonal. We assume them to be also normalized. Thus the c 's are simply the coefficients of expansion, determined directly by multiplying by the corresponding normal function and integrating. We must be careful of only one thing: our functions are now often complex, and when we multiply two such functions together, in such cases, it proves to be necessary always to multiply so that a function and a conjugate appear together. Thus we have

$$\iiint \psi(x, y, z, 0) \bar{u}_m(x, y, z) dx dy dz = \sum_n c_n \int u_n \bar{u}_m dx dy dz.$$

But now the orthogonality is such that $\int u_n \bar{u}_m dx dy dz$ is unity if $n = m$, zero if $n \neq m$, so that we have

$$c_m = \int \psi \bar{u}_m dv. \quad (9)$$

The physical situation is then this. If we know initially the distribution of coordinates, we can find a ψ satisfying the conditions, and in general all the c 's will be different from zero. That is, all overtones will be excited, or the system will be partly in each stationary state. We may say, if we choose, that we have an ensemble, and that a system of this ensemble has a probability $\bar{c}_n c_n$ of being in the n th state. If now we ask how ψ changes with time, we can see that the particle will no longer have the initial distribution of probability, but that the probability will change with time. For example, if we originally know it is in a small volume, this will not continue to be true as time goes on; it will have a chance of moving out of the volume. The reason is that the different waves cooperate to give just the right function at $t = 0$, but they vibrate with different frequencies, and soon they get out of step, and can no longer cooperate properly. Thus a general wave function, made by superposing many stationary states, does not represent an ensemble independent of time, though a single wave function does. Though the probabilities as functions of the coordinates change with time, it is significant that the c 's, being constants, do not. Thus the probability of finding the atom in a given stationary state does not change. The atoms do not go from one to another, and the states are really stationary. This is all true only

if we neglect radiation, or external forces. If there is radiation, the whole situation will be altered, the c 's will change with time, and the time rate of change of any $\bar{c}c$ will be interpreted as being connected with a corresponding probability that atoms are having transitions to or from this state. It is much as with vibrating strings: if the string is started off with a complicated shape, this shape will be soon changed, but if there is no friction we can analyze the motion into overtones, and each overtone preserves its amplitude. If friction is present, however, the overtones change their amplitudes.

227. Mean Values of Functions of Momenta.—The method of finding mean values of functions of the coordinates is perfectly straightforward, but the treatment of the momenta is peculiar, and is one of the characteristic features of wave mechanics. The momentum shows itself in the wave function through the wave length of the wave, and in order to get information about wave length, it turns out that the proper procedure is to differentiate the wave function. We can find the correct formulas from a very simple case; and since we are setting up a theory which is not derived from any other, we can do nothing but postulate the general formulas, which prove to be the same ones that we find in this special case. Thus suppose we have a free particle in empty space, traveling with a momentum p , energy E . Its wave function, if it travels along the x axis, will be $e^{\frac{2\pi i}{h}(px - Et)}$, corresponding to the wave length $1/\lambda = p/h$. More generally, if its components of momentum along the three axes are p_x , p_y , p_z , its wave function will be

$$e^{\frac{2\pi i}{h}(p_x x + p_y y + p_z z - Et)}, \quad (10)$$

a plane wave. If we wanted to find the mean x momentum of this particle, we should multiply p_x by the probability, and integrate; we should get p_x , of course, since the mean value of a constant is the same constant. But the question is, how is this to be generalized so that it can be used in more complicated cases, where the momentum does not appear explicitly, and is not constant? The answer proves to be the following: If our function is ψ , we observe that $\frac{h}{2\pi i} \frac{\partial \psi}{\partial x}$ equals $p_x \psi$. Thus if we form the expression $\int \frac{h}{2\pi i} \frac{\partial \psi}{\partial x}$ and integrate, the answer will be the same as integrating $\int p_x \psi$, which gives p_x . Similarly, we see that integrating $\int \left(\frac{h}{2\pi i} \frac{\partial}{\partial x} \right)^2 \psi$

would give p_x^2 , and so on. In other words, the operator $\frac{h}{2\pi i} \frac{\partial}{\partial x}$, operating on ψ , and averaged, can be taken to stand for the x component of momentum.

It is now assumed that this process can be applied in general. Thus with any wave function ψ , the mean value of the x component of momentum is $\int \psi \frac{h}{2\pi i} \frac{\partial}{\partial x} \psi dv$. Or more generally, if we have any function of momenta and coordinates, say $F(x, y, z, p_x, p_y, p_z)$, we have for the mean value

$$\bar{F} = \int \psi F\left(x, y, z, \frac{h}{2\pi i} \frac{\partial}{\partial x}, \frac{h}{2\pi i} \frac{\partial}{\partial y}, \frac{h}{2\pi i} \frac{\partial}{\partial z}\right) \psi dv. \quad (11)$$

This is the general rule, reducing to our former one when F involves only coordinates. There is one difficulty connected with this, however. It turns out that if there are any terms in F involving products of coordinates and momenta, the answer will depend on the order in which they occur. The best example is the case of the product $p_x x$. We have

$$\begin{aligned} \overline{p_x x} &= \int \psi \left[\frac{h}{2\pi i} \frac{\partial}{\partial x} (x\psi) \right] dv \\ &= \int \psi \left(\frac{h}{2\pi i} \psi + x \frac{h}{2\pi i} \frac{\partial \psi}{\partial x} \right) dv \\ &= \frac{h}{2\pi i} + \int \psi \left(x \frac{h}{2\pi i} \frac{\partial}{\partial x} \right) \psi dv \\ &= \frac{h}{2\pi i} + \overline{x p_x}, \end{aligned}$$

or

$$\overline{p_x x} - \overline{x p_x} = \frac{h}{2\pi i}. \quad (12)$$

This is the so-called commutation rule; it states that interchange, or commutation, of the order of a coordinate and momentum operator changes the value, since the difference is not zero. In most actual cases that we meet, we shall not be troubled by this difficulty of noncommutability of coordinates and momenta, but it is something against which we must be on our guard.

We notice by analogy with what we have done that, taking the wave function of the form given above, $-\frac{h}{2\pi i} \frac{\partial \psi}{\partial t} = E\psi$. This

again is taken to be a general method of finding the energy of a wave function:

$$\bar{E} = \int \bar{\psi} \left(-\frac{\hbar}{2\pi i} \frac{\partial}{\partial t} \right) \psi \, dv.$$

If $\psi = \sum_m c_m e^{-\frac{2\pi i E_m t}{\hbar}} u_m(x, y, z)$, we evidently have

$$-\frac{\hbar}{2\pi i} \frac{\partial \psi}{\partial t} = \sum_m c_m E_m e^{-\frac{2\pi i E_m t}{\hbar}} u_m(x, y, z).$$

Multiplying by $\bar{\psi}$, we have

$$\sum_{n, m} \bar{c}_n c_m E_m e^{-\frac{2\pi i (E_n - E_m)t}{\hbar}} \bar{u}_n u_m.$$

Integrating over the coordinates, the nondiagonal terms drop out on account of orthogonality of the u 's, and the rest reduces to

$$\bar{E} = \sum_n \bar{c}_n c_n E_n, \quad (13)$$

a weighted mean of the energy of the various states.

228. Schrödinger's Equation Including the Time.—We are now able to give a more general interpretation of Schrödinger's equation than was possible in Chap. XXIX. We start with the classical expression

$$H(q_1 \cdots q_n, p_1 \cdots p_n) = E,$$

where H is the Hamiltonian function, E is the total energy, and the equation represents the conservation of energy. But now suppose we try to replace each side by the corresponding quantum theory expression, so that we shall be able to allow each side to act on ψ , and if we wish multiply by $\bar{\psi}$ and integrate to get averages. The first step is

$$H\left(q_1 \cdots q_n, \frac{\hbar}{2\pi i} \frac{\partial}{\partial q_1}, \frac{\hbar}{2\pi i} \frac{\partial}{\partial q_2} \cdots \frac{\hbar}{2\pi i} \frac{\partial}{\partial q_n}\right) \psi = -\frac{\hbar}{2\pi i} \frac{\partial \psi}{\partial t}. \quad (14)$$

But this is just Schrödinger's equation, in the form involving the time (which we have not so far met). To show that it reduces to the form we have previously met, let us take the case of rectangular coordinates x, y, z . There

$$H = \frac{1}{2m}(p_x^2 + p_y^2 + p_z^2) + V,$$

so that the equation becomes

$$\left[-\frac{\hbar^2}{8\pi^2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) + V \right] \psi = -\frac{\hbar}{2\pi i} \frac{\partial \psi}{\partial t}.$$

In this, let us assume a solution $\psi = e^{-\frac{2\pi i}{\hbar}Et} u(x, y, z)$, where E is a constant, to be identified at the proper time with the energy. Then the equation becomes

$$\left(-\frac{\hbar^2}{8\pi^2m} \nabla^2 + V \right) u = Eu, \quad (15)$$

which leads immediately to the form of Schrödinger's equation with which we are familiar.

229. Some Theorems Regarding Matrices.—Suppose that we have an operator F , formed from a function of the q 's and p 's by replacing the p 's by differentiations, in the manner we have described. Then we have by definition $F_{nm} = \int \bar{u}_n F u_m dv$. But we can look at this in the following way. The u_n 's form a set of orthogonal unit vectors in function space. $F u_m$ is a function different in general from any of the u 's, and hence a different vector. The quantity $\int \bar{u}_n F u_m dv$ is the scalar product of $F u_m$ with \bar{u}_n ; that is, it is the component of $F u_m$ along the n th axis. But this suggests writing a vector equation:

$$F u_m = \sum_n F_{nm} u_n, \quad (16)$$

expressing $F u_m$ as a sum of unit vectors, each multiplied by the corresponding component. To prove this, we need only multiply by \bar{u}_n and integrate, when the right side, on account of orthogonality, leaves only F_{nm} . An example of such an expression is Schrödinger's equation not involving the time, which can be written

$$H u_n = E_n u_n, \quad (17)$$

if E_n is the energy in the n th state. This obviously expresses the fact that the matrix of H has only diagonal components ($H_{nm} = E_n$ if $n = m$, zero if $n \neq m$), so that, since H has no nondiagonal components, it has no terms depending on time, or is a constant.

It is interesting to write down the matrix of a constant, for example a number C . Evidently $C_{nm} = \int \bar{u}_n C u_m dv = C$ if $n = m$, 0 if $n \neq m$. A particular case is the matrix of unity, $\int \bar{u}_n u_m dv = 1$ if $n = m$, 0 if $n \neq m$, simply the orthogonality and normalization conditions. This matrix is often called δ_{nm} ; by

definition $\delta_{nm} = 1$ if $n = m$, 0 if $n \neq m$. In terms of this, we have $C_{nm} = C\delta_{nm}$. And we can write the matrix of the energy as

$$H_{nm} = E_n \delta_{nm}. \quad (18)$$

This matrix equation, stating that the matrix of the energy is a diagonal matrix with the characteristic values E_n , may be taken as a matrix statement of Schrödinger's equation; we readily see that it is just what would be obtained by multiplying Schrödinger's equation by an arbitrary u_m , and integrating. We shall actually use this matrix equation later in discussing perturbation theory. It is to be noted that a matrix depends on two things: first, the operator, and secondly, the set of orthogonal functions with respect to which it is computed. Thus a given operator, as energy or angular momentum or x coordinate, can have its matrix computed in any set of orthogonal functions. The problem of solving Schrödinger's equation with a given energy operator may be considered as that of finding the particular set of orthogonal functions which makes that operator diagonal. In a similar way we can find a set of orthogonal functions which would make any other desired operator have a diagonal matrix. We shall see in the next chapter that this involves us essentially in a rotation of axes in function space, similar to what we found in introducing normal coordinates in vibration problems.

From our expansion of Fu_m in series in the u_n 's, we can easily get the method of multiplying matrices, which is very useful in matrix manipulation. Suppose that we have two operators F and G , and know the matrix components F_{nm} and G_{nm} . We can then find easily the matrix components of the product operator FG . For we have

$$\begin{aligned} Gu_n &= \sum_m G_{mn} u_m \\ FG u_n &= \sum_m G_{mn} F u_m = \sum_{m,k} G_{mn} F_{km} u_k = \sum_k \left(\sum_m F_{km} G_{mn} \right) u_k. \end{aligned}$$

But also

$$(FG)u_n = \sum_k (FG)_{kn} u_k,$$

by the earlier formula. Hence

$$(FG)_{kn} = \sum_m F_{km} G_{mn}, \quad (19)$$

the formula for multiplying matrices.

It is a rather remarkable fact that the method of operating with matrices was discovered before the wave mechanics. This multiplication rule, and the commutation rule, were both developed. They were used for a number of complicated calculations, without use of wave functions, for example for finding the energy levels of the linear oscillator, its intensities of radiation, and even the energy levels of the hydrogen atom. For a few problems, as perturbation theory, the matrix method is still more convenient than the wave method, as we shall see.

Problems

1. Prove that a coordinate commutes with another coordinate; a momentum commutes with another momentum; and a coordinate commutes with a momentum conjugate to another coordinate.

2. Write down the operators for the three components of angular momentum in rectangular coordinates.

3. If F is any operator, prove that $\frac{\hbar}{2\pi i} \frac{\partial F}{\partial t} = (HF - FH)$, where H is the Hamiltonian operator, the equation above to be regarded as either an operator or matrix equation. To prove it, take average values of the operators. Find the average value of F , differentiate it with respect to time, to get the left side of the expression. On the right, in computing the average values, use the multiplication rule to compute the matrices of HF and FH , noting that H has a diagonal matrix. Finally identify terms on both sides of the equation.

4. Using the result of Prob. 3, prove that the time rate of change of the energy is zero; prove that H and t satisfy a commutation relation like a momentum and coordinate.

5. Show that for the linear oscillator the assumptions

$$\begin{aligned} E_n &= (n + \frac{1}{2})\hbar\nu \\ x_{nn} &= 0 \\ x_{n+1,n} &= x_{n,n+1} = \sqrt{\frac{n(n+1)}{8\pi^2 m \nu}} \\ x_{nm} &= 0 \text{ if } m \neq n \pm 1 \end{aligned}$$

satisfy the quantum mechanics. To do this, compute the matrix components of \hat{x}_{nm} , and find the matrix of the energy expression $(m/2)(\dot{x}^2 + 4\pi^2\nu^2 x^2)$, computing the matrices of \dot{x}^2 and x^2 by the multiplication rule. Show that this matrix is diagonal, its diagonal components being the energy values given above.

6. By comparing with the wave functions of the linear oscillator in Chap. XXIV, Prob. 6, verify that the values of matrix components in Prob. 5 are correct. If you cannot give a general proof, take the actual wave functions you have worked out, in Prob. 6, Chap. XXIX, using them for $n = 0, 1, 2$, normalizing, and calculating the matrix components by direct integration.

7. Show that a linear oscillator radiating from the n th stationary state cannot jump except to the $(n - 1)$ st state, so that there is a selection principle on its radiation. Compute the rate of radiation of the oscillator in the n th state, on the assumption that it is the same as that of a classical

oscillator whose charge is e , displacement is $x_{n,n-1}e^{\frac{2\pi i}{h}(E_n - E_{n-1})t} + x_{n-1,n}e^{\frac{2\pi i}{h}(E_{n-1} - E_n)t}$. Compare this displacement with the displacement of a classical oscillator of energy E_n , showing that in the limit of large quantum numbers both amplitude and frequency of the classical oscillator agree with the quantum values. This is an example of the correspondence principle.

8. Solve Schrödinger's equation for a rotator, whose kinetic energy is $\frac{1}{2}I\dot{\theta}^2$, in the absence of an external force. Find wave functions, showing that the angular momentum is an integral multiple of $\hbar/2\pi$. Compute the radiation, by finding matrix components of $x = r \cos \theta$, $y = r \sin \theta$, which determine the displacement, and show that the only allowed transitions are those in which the angular momentum changes by ± 1 unit.

9. Find what $p^2q - qp^2$ is equal to, using the commutation rule for $pq - qp$.

10. Show that $e^{\frac{2\pi i}{h}p\alpha}u(x)$, where p is the x component of momentum, α is a constant, is equal to $u(x + \alpha)$. Use Taylor's expansion of the exponential operator.

11. Write down Schrödinger's equation in spherical polar coordinates, by using the Laplacian in these coordinates, assuming a potential $V(r)$. Discuss the method of deriving the equation from the Hamiltonian by replacing the momenta by differentiations, showing that the former method is consistent with the latter, but that the latter method does not lead to unique results.

CHAPTER XXXII

PERTURBATION THEORY

There are many problems in wave mechanics, which, though they cannot be exactly solved, are approximated by soluble problems. Thus a nonlinear oscillator can be approximated by a linear one; or a system, as an atom, in an external electric or magnetic field can be approximated by the same system without the field. The perturbation theory is adapted to the solution of such problems, starting with the known approximate solution, and expanding in power series in the perturbation. At the same time, there are some problems of more general nature treated by perturbation theory. Thus the radiation of an atom can be examined by treating the interaction of the atom and a radiation field as a perturbation. We shall be led by such questions to a discussion of the transitions between stationary states. The actual method we shall use is closely analogous to the perturbation theory used with the nonuniform vibrating string.

230. The Secular Equation of Perturbation Theory.—Suppose that we wish to solve Schrödinger's equation $Hu_n = E_n u_n$, where H is the given Hamiltonian. Let us start with a set of orthogonal functions u_n^0 , which often are solutions of a similar problem approximating the real one, and let us expand the correct functions u_n in series in the u_n^0 's:

$$u_n = \sum_m S_{mn} u_m^0. \quad (1)$$

Then the problem may be regarded as that of finding the expansion coefficients S_{mn} , which are really coefficients of a linear transformation in function space transforming from the original set of orthogonal functions to the final, correct, ones, so that we may expect the S 's to satisfy orthogonality and normalization conditions. We substitute this expression for u_n in Schrödinger's equation, and get the condition for the coefficients. If we substitute, multiply by \bar{u}_k^0 , and integrate, we shall have only

one term on the right, on account of orthogonality of the u^0 's; on the left, we shall have a linear combination, each term involving a matrix component of H with respect to the u^0 's, for example $H_{km} = \int \bar{u}_k^0 H u_m^0 dv$. We recall that, since the u^0 's are not solutions of the problem, this matrix will not be diagonal. Carrying out the substitution, we have

$$\sum_m (H_{km} - E_n \delta_{km}) S_{mn} = 0, \quad (2)$$

or an infinite set of equations for the infinite set of S_{mn} 's. Writing them for the n th stationary state, we have

$$\begin{aligned} &(H_{11}-E_n)S_{1n}+H_{12}S_{2n}+H_{13}S_{3n}+\cdots=0 \quad (k=1) \\ &H_{21}S_{1n}+(H_{22}-E_n)S_{2n}+H_{23}S_{3n}+\cdots=0 \quad (k=2) \\ &\qquad\vdots \end{aligned}\tag{3}$$

These equations are all homogeneous, of the same sort found whenever we have introduced normal coordinates or rotated axes, as, for example, in discussing coupled systems or the vibrating string. As usual, the equations in general do not have a solution; they have one only if the determinant of coefficients

$$\begin{vmatrix} H_{11} - E_n & H_{12} & H_{13} & \dots \\ H_{21} & H_{22} - E_n & H_{23} & \dots \\ \dots & \dots & \dots & \dots \end{vmatrix} \quad (4)$$

is zero. This secular equation determines the energy levels.

231. The Power Series Solution.—If the u^0 's were solutions of the problem, H would have a diagonal matrix, the diagonal terms being the energy levels. Though this is not true, let us assume that the u^0 's are not far from solutions. Then by arguments of continuity the nondiagonal terms of H , though not zero, are small, and the diagonal terms, though not exactly the energy values E_n of the exact solution, are not far from the correct values. Thus E_n is approximately H_{nn} . We assume the problem is nondegenerate, by which we mean that only one state has even approximately this same value. Now let us recall how to expand a determinant. We take products of terms, choosing just one from each row, one from each column. There are $N!$ ways of doing this, if the determinant has N rows and columns. We give each a sign $+$ or $-$ according to its requiring an even or an odd number of interchanges of rows or columns to bring the desired term to the principal diagonal. Finally we add. In this case, since we are dealing with small

quantities, we look first for the largest product. This is plainly the principal diagonal, for the only large terms are those on the principal diagonal. For a first approximation we may set this equal to zero. It is already factored: $(H_{11} - E_n)(H_{22} - E_n) \cdots = 0$. One of the factors must be, to this approximation, zero. Plainly it must be $H_{nn} - E_n$, since this is the only term which is even small, assuming the system is nondegenerate. This then is the first order approximation to the energy: $E_n = H_{nn}$, the diagonal component of the matrix of the energy with respect to the approximate wave function.

Using the first-order approximation to the energy, we can easily get the corresponding linear transformation and wave functions. If the u^0 's were the correct wave functions, we should have $S_{nn} = 1$, all the other S 's = 0. To a first approximation, in the actual case, we may set $S_{nn} = 1$, but regard the other S 's as small quantities. Then we have, for example, for the first equation

$$(H_{11} - H_{nn})S_{1n} + \cdots + H_{1n} + \cdots = 0,$$

where the terms we do not write are of a smaller order than those we write. Hence

$$S_{1n} = -\frac{H_{1n}}{H_{11} - H_{nn}}. \quad (5)$$

The other equations are of the same form, so that the approximate wave function is

$$u_n = u_n^0 - \sum_{k \neq n} \frac{H_{kn}u_k^0}{H_{kk} - H_{nn}}. \quad (6)$$

For the second approximation to the energy, we must consider further terms in the determinant. We can proceed by analogy with the case of a determinant of two rows and two columns, which we should have if there were only two stationary states to consider. In this case the secular equation would be

$$\begin{vmatrix} H_{11} - E_n & H_{12} \\ H_{21} & H_{22} - E_n \end{vmatrix} = (H_{11} - E_n)(H_{22} - E_n) - H_{12}H_{21} = 0. \quad (7)$$

This is only a quadratic for E_n , and can be immediately solved explicitly:

$$\begin{aligned} E_n &= \frac{H_{11} + H_{22}}{2} \pm \sqrt{\left(\frac{H_{11} + H_{22}}{2}\right)^2 - (H_{11}H_{22} - H_{12}H_{21})} \\ &= \frac{H_{11} + H_{22}}{2} \pm \sqrt{\left(\frac{H_{11} - H_{22}}{2}\right)^2 + H_{12}H_{21}}. \end{aligned} \quad (8)$$

This explicit formula is analogous to the formula for the frequency of a system of two coupled oscillators, obtained in Eq. 4, Chap. XI. Here as there, if the nondiagonal matrix component H_{12} of the energy is small, we can expand the radical by the binomial theorem, obtaining without trouble for the two solutions as power series in H_{12} ,

$$\begin{aligned} E_1 &= H_{11} + \frac{H_{12}H_{21}}{H_{11} - H_{22}} + \dots \\ E_2 &= H_{22} + \frac{H_{12}H_{21}}{H_{22} - H_{11}} + \dots \end{aligned} \quad (9)$$

analogous to Eq. (5) of Chap. XI. Here, as there, the effect of the second-order perturbation terms is to push apart the two levels. Thus the first-order calculation alone gives $E_1 = H_{11}$. The numerator of the fraction giving the second-order calculation, $H_{12}H_{21}$, is really a perfect square, for it can be shown that $H_{21} = H_{12}$ (similar theorems hold in general for all the matrices of real quantities which we meet). Thus the numerator is positive. If H_{11} is greater than H_{22} , so that the first-order level 1 lies above 2, the denominator is also positive, so that the level is still further raised by this perturbation. On the other hand, for the other level, the denominator is negative, and the level is further depressed.

The exact solution which we have obtained in Eq. (8) is only possible when the secular equation is simple enough to handle algebraically. The approximations (9), however, can be found directly from the secular equation (7). Thus let us consider E_1 . We assume that the equation is not degenerate, so that $H_{22} - E_1$ is not a small quantity, and we may divide Eq. (7) by it. Thus we have

$$H_{11} - E_1 = \frac{H_{12}H_{21}}{H_{22} - E_1}.$$

Replacing the E_1 in the denominator by its value H_{11} , which is correct to the first order, this becomes

$$E_1 = H_{11} + \frac{H_{12}H_{21}}{H_{11} - H_{22}},$$

agreeing with Eq. (9). By a little consideration of the determinant, exactly a similar discussion can be given in the general case. And the result proves to be simply

$$E_n = H_{nn} - \sum_{k \neq n} \frac{H_{kn}H_{nk}}{H_{kk} - H_{nn}} \dots, \quad (10)$$

in agreement with the special case solved above. It is very rarely that further approximations than we have given are used, for either the energy or the wave function.

232. Perturbation Theory for Degenerate Systems.—We shall often meet cases in which the unperturbed problem is degenerate; that is, where the diagonal energies H_{nn} of several states are almost exactly equal to each other. In this case, the power series method evidently does not work; the differences of energy which appear in the denominator of the terms in Eq. (9) or (10) become zero, or very small, and the series diverge and even have infinite terms. If there were only two levels, as in the special case taken up in the last section, we could solve the problem explicitly, not using the power series at all. Thus if $H_{11} - H_{22} = 0$, Eq. (8) gives

$$E = H_{11} \pm H_{12}, \quad (11)$$

an important formula for perturbations of degenerate systems. With a finite number of degenerate levels, we have a secular equation of finite degree, and while we cannot solve it as conveniently as the quadratic, still we can approximate its solutions, even for the degenerate case where the differences of diagonal energies are smaller than the nondiagonal energy terms. Now it fortunately happens that in many problems in which degeneracy enters, as in atomic spectra, the levels fall into groups, the energies of all the levels in a group being about the same, but the different groups being well separated in energy. Such groups of levels are the multiplets in atomic spectra. In these cases we first solve the problem of the levels within a group, finding an exact solution for the finite secular equation. This solution gives us not only energy levels, but also coefficients of linear combinations transforming the original wave functions of the group into a new set which has the property that it makes the matrix components of the energy diagonal, with respect to the states of this group. We then use these transformed functions as the starting point of a new perturbation calculation, in which perturbations between adjacent groups are considered. In terms of these transformed functions, the energy will have no nondiagonal components between levels which lie close to each other, in the groups, but only between levels in different groups, at a considerable energy distance apart. Thus we may use the series method of Eq. (10), and the second-order terms will

be small, since the only terms of the summation for which the denominator is small will have numerators equal to zero. It is to be particularly noted in this discussion that the difficulty in applying the power series method to degenerate systems arises, not on account of any unusual size of the nondiagonal energy components, but on account of the unusually small energy differences between diagonal terms. The method converges only if the nondiagonal component between any two levels is small compared with the difference of diagonal energies of the two terms. This demands that before applying the power series method the nondiagonal terms between degenerate levels be removed, but it imposes no such requirement on the terms between levels of quite different energy.

We can see more clearly what is happening from a mechanical analogy. Suppose we have a large number of mechanical oscillators coupled together, all having different natural frequencies, except the first two, which have unperturbed frequencies exactly, or almost exactly, the same. In considering the interaction, the effect of the two of equal frequency on each other will be large, since each one resonates with the other, but the others will have much less effect. We, therefore, first solve only the interaction of these two resonating oscillators, introducing normal coordinates for them. Then we can proceed with the discussion of the interaction, treating the effect of the other oscillators, not on these two oscillators individually, but on the two normal coordinates representing them. Of course, if there are several groups of degenerate levels, we introduce changes of variables inside each group first, then apply the ordinary perturbation theory. We shall have many examples of degenerate systems in our discussion of atomic structure, where nearly every energy level of an unperturbed atom is degenerate, and is split up by an external perturbing field, as an electric or magnetic field. In more complicated atoms, the perturbing fields come from within the atom itself, being interactions of one part on another, producing the multiplet structure. In actual practice, we shall find the study of degenerate systems very important.

233. The Method of Variation of Constants.—A slightly different point of view in perturbations is obtained by considering the time variation. Let us expand ψ , the correct wave function depending on time, in series in the unperturbed functions

$u^0: \psi = \sum_m C_m(t) u_m^0(x)$, where the C 's—functions of time—

would be pure exponentials, $c_n e^{-\frac{2\pi i}{h} E_n t}$, if the u^0 's were the correct solutions of the problem. Whether correct or not, we can always make the expansion above, for at any instant ψ can be expressed in series in the orthogonal functions u^0 , the coefficients being functions of time. Now let us try to satisfy the equation

$$H\psi = -\frac{h}{2\pi i} \frac{\partial \psi}{\partial t}. \quad \text{We have}$$

$$\sum_m \left(C_m H u_m^0 + \frac{h}{2\pi i} \frac{dC_m}{dt} u_m^0 \right) = 0.$$

Multiplying by \bar{u}_k^0 and integrating, the result is

$$\frac{dC_k}{dt} = -\frac{2\pi i}{h} \sum_m H_{km} C_m. \quad (12)$$

These equations for the time derivatives of the C 's in terms of their instantaneous values are enough to determine the complete solution of the problem.

To make connection with the ordinary method, we need only assume $C_m = S_{mn} e^{-\frac{2\pi i}{h} E_n t}$, an exponential solution. Then immediately we have, canceling the exponential, and the factor $-2\pi i/h$,

$$E_n S_{kn} = \sum_m H_{km} S_{mn},$$

or exactly the equation we have previously used. In more general cases, however, it is not always possible to make this assumption. An example is that in which the perturbative force depends on the time.

234. External Radiation Field.—The most interesting example of the method of variation of constants is the perturbation by an external radiation field, for this actually produces transitions between stationary states. First let us look a moment at the physical side of the problem, so as to understand what we expect to obtain from the calculations. An ordinary radiation field is never exactly sinusoidal; its amplitude, at a given point of space, as function of time, may be analyzed in Fourier series of very long period, as in Sec. 185, Chap. XXV. If the field is approximately monochromatic of frequency ν_0 , that means that only

frequencies in the neighborhood of ν_0 will have large amplitudes in the Fourier representation. On the other hand, if it is continuous radiation, as the radiation from hot solids, there will be considerable amplitude in all frequencies, at least over a certain region. We assume the latter case. The electric field in the x direction at a given point will then be $\sum_{\nu} E_{\nu} \cos 2\pi(\nu t - \alpha_{\nu})$,

where E_{ν} , α_{ν} , are amplitude and phase of the component of frequency ν , and where we have components of frequencies differing by small increments $d\nu = 1/T$, where T is the fundamental period. The phases α_{ν} of successive components may be treated as being statistically independent of each other; that is, if we take any two components, the chance that the phase angle between them at any instant should have one value is just equal to the chance that it have another value. The values of E_{ν} will be treated as functions of ν , though a somewhat more general treatment subjects them to probability laws too. Now we are interested in finding $\rho_{\nu} d\nu$, the energy per unit volume in the frequency range $d\nu$. Since one component of the series is associated with the range $d\nu = 1/T$, we can simply find the energy of this component. For the x component of electric field, this is $\frac{1}{8\pi}[2E_{\nu}^2 \cos^2 2\pi(\nu t - \alpha_{\nu})]$, the factor 2 taking account of the magnetic field as well as the electric field. The time average of this term is $E_{\nu}^2/(8\pi)$. If we are dealing with radiation having equal intensities in all directions, the mean energy per unit volume associated with x , y , and z coordinates will be equal. Hence we have

$$\rho_{\nu} d\nu = \frac{3}{8\pi} E_{\nu}^2. \quad (13)$$

235. Einstein's Probability Coefficients.—Now suppose a radiation field of the type we have described is allowed to act on an atomic system. Einstein was the first to solve this problem. He assumed that, if the atom is in its m th state, there will be the following probabilities of transition to other states, induced by the radiation field:

1. A probability A_{mn} of radiating spontaneously to each state n which is of lower energy than the m th, with emission of the corresponding photon of frequency ν_{mn} , given by $E_m - E_n = h\nu_{mn}$. This spontaneous emission corresponds to the ordinary

radiation of an oscillating dipole in classical electromagnetic theory.

2. A probability $B_{mn}\rho_{mn}$ of absorbing a photon of frequency ν_{mn} from the radiation field, where now the state n has higher energy than m , and of jumping up to the state n . This probability is proportional to the energy density ρ_{mn} at the particular frequency ν_{mn} in the external radiation field.

3. A probability $B_{mn}\rho_{mn}$, where now the n th state lies below the m th, of emitting a photon of frequency ν_{mn} , and falling to the lower state, under action of the radiation. This is called induced or forced emission.

Einstein assumed that the following relations held between the A 's and B 's corresponding to any transition $n - m$, where $E_m > E_n$: $B_{mn} = B_{nm}$, and $A_{mn}/B_{mn} = 8\pi h \nu^3_{mn}/c^3$. Assuming these simple laws, he could then give a very elementary derivation of Planck's law of black-body radiation. Let us assume that we have a piece of matter containing many kinds of atoms, so as to have some capable of emitting and absorbing each frequency. Consider a particular set of atoms having a lower state 1, an upper state 2, and assume that at temperature T the number of atoms in the upper state is to the number in the lower state as $e^{-E_2/kT}$ is to $e^{-E_1/kT}$, or the Maxwell-Boltzmann distribution law. Now we ask, what intensity, or energy density, in the external radiation field must we have to be in equilibrium with these atoms? If we can find this for each frequency of radiation, we shall necessarily have the distribution of intensity in radiation in equilibrium with matter at temperature T , which is what Planck's law gives. Let N_2 be the number of atoms in the second state, N_1 in the first, so that $N_2/N_1 = e^{-(E_2-E_1)/kT} = e^{-h\nu/kT}$, where ν is the frequency emitted or absorbed by the atom in its transition. Now we know that the number of atoms leaving the second state per second is equal to the sum of the following:

1. The number leaving on account of spontaneous radiation, or $N_2 A_{21}$.

2. The number entering on account of absorption from the lower state, or $-N_1 B_{12}\rho_{12}$.

3. The number leaving on account of induced emission, or $N_2 B_{21}\rho_{12}$. This sum must be zero, in a steady state where the N 's are constant. Hence

$$N_2(A_{21} + B_{21}\rho_{12}) = N_1 B_{12}\rho_{12}.$$

Using the relation between the A 's and B 's, this is

$$N_2 B_{12} \left(\rho_{12} + \frac{8\pi h \nu^3}{c^3} \right) = N_1 B_{12} \rho_{12}.$$

Setting $N_2/N_1 = e^{-h\nu/kT}$, canceling B_{12} , and solving for ρ_{12} , we have

$$\rho_{12} = \frac{8\pi h \nu^3}{c^3} \frac{1}{e^{h\nu/kT} - 1}, \quad (14)$$

which is Planck's law of black-body radiation.

236. Method of Deriving the Probability Coefficients.—Einstein's coefficient A is often derived by analogy with classical theory as follows: In Chap. XXXI we have seen that the matrix components of electric moment are connected with probabilities of radiation. Thus, if the amplitude of the component of moment of the atom corresponding to the transition 2—1 is C , the corresponding classical rate of radiation is $\frac{16\pi^4 C^2 \nu^4}{3c^3}$. We can write this component in terms of the matrices as follows: corresponding to this frequency, we have the terms $(ex)_{12} e^{-\frac{2\pi i}{h}(E_2 - E_1)t} + (ex)_{21} e^{-\frac{2\pi i}{h}(E_1 - E_2)t} = 2(ex)_{12} \cos 2\pi \nu t$, where $h\nu = E_2 - E_1$. Thus $C = 2(ex)_{12}$, and the rate of radiation is $\frac{64\pi^4 (ex)_{12}^2 \nu^4}{3c^3}$. But an atom with a probability A_{21} of radiating a photon of energy $h\nu$ is radiating on the average at the rate of $A_{21}h\nu$ per second. Hence we must set this equal to the rate of radiation above, giving

$$A_{21} = \frac{64\pi^4 (ex)_{12}^2 \nu^3}{3c^3 h}, \quad B_{21} = A_{21} \frac{c^3}{8\pi h \nu^3} = \frac{8\pi^3 (ex)_{12}^2}{3h^2}. \quad (15)$$

The argument given above is hardly a derivation; it is merely suggestive. To get a real derivation of the probabilities, we use the method of perturbations. We shall find, for a reason to be discussed in a later section, that we can only obtain the B 's by this method. We shall assume that at $t = 0$ the atom is definitely in the m th state; that is, $c_m^0 = 1$, all other c 's are zero, where the c 's are the coefficients in the expansion of the wave function ψ in terms of the unperturbed stationary states, so that $\bar{c}_n c_n$ is the probability of finding the system in the n th state, and the c 's are the values when $t = 0$. Then we shall investigate the time variation of the c 's by the method of variation of constants, and it will appear that the c 's for n different from m increase

linearly with time, so long as we consider only small intervals of time and small perturbations, the term $\bar{c}_m c_m$ correspondingly decreasing. This we interpret as a definite probability that the system will leave the m th state and go to the n th; in fact, we shall find $\bar{c}_n c_n$ equal exactly to $B_{mn} \rho_{mn} t$, as far as the variation is linear with time. By comparing this expression with the derived values of $\bar{c}_n c_n$, we can evaluate the B 's directly from perturbation theory.

237. Application of Perturbation Theory.—Let the Hamiltonian of the system without radiation be H^0 , and assume that the unperturbed problem can be solved exactly:

$$H^0 u_m^0 = E_m^0 u_m^0.$$

Let the perturbed Hamiltonian be $H^0 - ex \sum_{\nu} E_{\nu} \cos 2\pi(\nu t - \alpha_{\nu})$,

the second term representing the potential of the force of the field represented by the summation, on the charge e . Under the action of the perturbation, let the perturbed wave function be $\psi = \sum_m C_m(t) u_m^0(x)$. Our task now is to find the C 's. Using the method of variation of constants, noting that H^0 has a diagonal matrix, we have

$$\begin{aligned} \frac{dC_n}{dt} &= -\frac{2\pi i}{h} \sum_k H_{nk} C_k \\ &= -\frac{2\pi i}{h} H_{nn}^0 C_n + \frac{2\pi i}{h} \sum_k (ex)_{nk} C_k \sum_{\nu} E_{\nu} \cos 2\pi(\nu t - \alpha_{\nu}). \end{aligned}$$

Now let $C_n = c_n(t) e^{-\frac{2\pi i}{h} H_{nn}^0 t}$, where $c_n(t)$ would be constant in the absence of an external field. Writing the field in exponential form, and letting $H_{nn}^0 - H_{kk}^0 = h\nu_{nk}$, this gives

$$\frac{dc_n}{dt} = \frac{2\pi i}{h} \sum_k (ex)_{nk} c_k \sum_{\nu} \frac{E_{\nu}}{2} \{ e^{2\pi i[(\nu + \nu_{nk})t - \alpha_{\nu}]} + e^{-2\pi i[(\nu - \nu_{nk})t - \alpha_{\nu}]} \}$$

If the external field were not present, we plainly would have $dc_n/dt = 0$; if there is a small field, the time derivative will be small, or, in other words, the c 's will be approximately constant. To a first approximation we may assume on the right side that the c 's are exactly constant, having the values c^0 which they had at $t = 0$. If this is so, we may integrate directly, obtaining

$$\begin{aligned}
c_n - c_n^0 &= \int_0^t \frac{dc_n}{dt} dt \\
&= \sum_{k, \nu} \frac{(ex)_{nk}}{2\hbar} c_k^0 E_\nu \left[e^{-2\pi i \alpha_\nu} \left(\frac{e^{2\pi i (\nu + \nu_{nk})t} - 1}{\nu + \nu_{nk}} \right) \right. \\
&\quad \left. - e^{2\pi i \alpha_\nu} \left(\frac{e^{-2\pi i (\nu - \nu_{nk})t} - 1}{\nu - \nu_{nk}} \right) \right]. \quad (16)
\end{aligned}$$

Now let us take the case we have discussed, where at $t = 0$ we have $c_m^0 = 1$, all the other c 's zero. Then for any $n \neq m$, we have only the single term of the summation above for which $k = m$. Next we find $\bar{c}_n c_n$. In this, we have a product of two sums over ν , which is, therefore, a double sum. Each such term for which we have different frequencies in the two factors has a term $e^{-2\pi i (\alpha_\nu - \alpha_{\nu'})}$, which, on account of the random nature of the phases, is as likely to be positive as negative, and on the average cancels. Thus we are left with only the squares of the individual terms, in which the α 's drop out. Further, each of these squares has terms whose denominators are respectively $(\nu + \nu_{nm})^2$, $(\nu + \nu_{nm})(\nu - \nu_{nm})$, and $(\nu - \nu_{nm})^2$. The frequency ν_{nm} is so defined that it is positive if the n th state lies above the m th, which we assume to be the case for the moment. When ν becomes nearly equal to ν_{nm} , the term $(\nu - \nu_{nm})^2$ is very small, the term with this as denominator very large. Since ν is always positive, it is not possible for the other terms, involving $\nu + \nu_{nm}$ in the denominator, to become so large. To an approximation, then, we neglect all terms except the last, obtaining

$$\begin{aligned}
\bar{c}_n c_n &= \frac{(ex)_{nm}^2}{4\hbar^2} \sum_\nu E_\nu^2 \frac{[e^{2\pi i (\nu - \nu_{nm})t} - 1][e^{-2\pi i (\nu - \nu_{nm})t} - 1]}{(\nu - \nu_{nm})^2} \\
&= \frac{(ex)_{nm}^2}{2\hbar^2} \sum_\nu E_\nu^2 \frac{[1 - \cos 2\pi (\nu - \nu_{nm})t]}{(\nu - \nu_{nm})^2} \\
&= \frac{(ex)_{nm}^2}{\hbar^2} \sum_\nu E_\nu^2 \frac{\sin^2 \pi (\nu - \nu_{nm})t}{(\nu - \nu_{nm})^2}. \quad (17)
\end{aligned}$$

The formula we have just derived is decidedly significant. It gives essentially the probability that the system will go, in time t , from state m to state n , under the action of the radiation. For a particular frequency ν , this probability is seen to be proportional to E_ν^2 ; that is, to the intensity of the incident radiation; and to $(ex)_{nm}^2$, the square of the matrix of the electric moment

connected with this particular transition, which we should expect. But in addition, there is a dependence on frequency. If we plot $\bar{c}_n c_n$, at time t , against ν , the impressed frequency, we get a narrow peak with small side bands, centering at ν_{nm} , just like the pattern found in Fraunhofer diffraction. Thus, if the impressed frequency is close to the absorption frequency ν_{nm} , there will be a large probability of transition, while if it is farther away, the probability will be smaller. If the perturbation acts only for a small time, the band will be broad, indicating that many frequencies can cause the transition, but if the time is long enough, practically only the frequency ν_{nm} can cause the transition; the absorption curve of the substance, in other words, will have a sharp absorption line corresponding to the various transitions from the state m to other states n , as calculated by the quantum theory.

In carrying out the summation over ν , it is evident that the essential contributions will come for frequency ν very close to ν_{nm} . In this region, we may replace E_ν by its value at ν_{nm} , which we have already seen to be given by $3E_{\nu_{nm}}^2/8\pi = \rho_{nm}d\nu$. Hence the summation reduces to an integration,

$$\bar{c}_n c_n = \frac{8\pi(ex)^2_{nm}}{3h^2} \rho_{nm} \int \frac{\sin^2 \pi(\nu - \nu_{nm})t}{(\nu - \nu_{nm})^2} d\nu. \quad (18)$$

The integration should properly be taken from $\nu = 0$ to infinity. But since the integrand is large only in the immediate neighborhood of $\nu = \nu_{nm}$, we shall make a negligible error if we integrate from $-\infty$ to ∞ . Then the integral becomes $\pi t \int_{-\infty}^{\infty} \frac{\sin^2 z}{z^2} dz$, where $z = \pi(\nu - \nu_{nm})t$. This can be easily evaluated, giving $\pi^2 t$. Thus we have finally

$$\bar{c}_n c_n = \frac{8\pi^3(ex)^2_{nm}}{3h^2} \rho_{nm} t, \quad (19)$$

or $B_{nm}\rho_{nm}t$, where B_{nm} is as given before. Thus we have verified our earlier statement regarding the probability coefficients B . A simple variation of the argument applies to states n of lower energy than the state m , resulting in the probability of forced emission, and if we compute $\bar{c}_m c_m$, we find that the number of systems in the m th state decreases at a rate to compensate the increase in the other states. This can be shown easily on general grounds as well as by direct computation, for it can be

shown that the sum of the quantities $\bar{c}_n c_n$ for all states remains constant.

238. Spontaneous Radiation and Coupled Systems.—The calculation we have just given did not lead to the probability of spontaneous emission A_{nm} . An attempt might be made to include it by adding to the external force a radiation resistance term, depending on the velocity of the electron, but this method proves not to lead to the right answer. The proper treatment, as a matter of fact, must be sought in a different direction. We treat the radiation field, not as a perturbation, but as part of the system. It is possible to apply the quantum theory directly to the field by itself. For instance, if the radiation is confined in a rectangular box with perfectly reflecting walls, the electromagnetic field inside consists of a set of standing waves, of all the wave lengths allowed for a vibrating solid of the corresponding size, and with corresponding frequencies. We can now introduce normal coordinates, each corresponding to one mode of vibration, and the classical equations of motion of these normal coordinates are just like those of a linear oscillator. In a corresponding way, in wave mechanics, we treat these normal coordinates, set up a wave equation for each, and find that each one is quantized, with energy $(n + \frac{1}{2})h\nu$, where ν is the frequency of the wave, n a quantum number associated with this particular mode of vibration. A change of this quantum number by unity corresponds to an increase or decrease of the energy of the radiation field by one unit $h\nu$, and this we identify with the creation or destruction of a photon of this energy, by interaction with matter.

Next we treat the atomic system just as if the radiation were not present. In this case, the atom will continuously stay in the same stationary state, and similarly the radiation field will always keep the same quantum numbers, meaning that no photons are being created or destroyed. But finally we introduce into the complete system of atoms and radiation a perturbation, corresponding to the potential of the atom in the radiation field (including the vector as well as scalar potential). This couples the two systems together, and under the influence of the perturbation transitions are possible, in which the atoms gain or lose energy in passing between stationary states, and the radiation field loses or gains an equal energy, which appears as destruction or creation of corresponding photons, or decrease

or increase of the quantum number of the proper normal vibration of the radiation. When the probability of these processes is investigated, by the method of variation of constants, it is found that we obtain not only the probability of forced absorption or emission, $B\rho$, but also the probability of spontaneous emission A . It is not hard by this method to investigate other questions as well, as for instance the breadths of absorption or emission lines—the question of just what frequencies of light can interact with a given atomic system. The general result is that, the shorter the life of an atom in either the upper or lower state associated with a transition, the broader the corresponding absorption or emission line.

It is interesting to look a little more closely at the sort of perturbation problem we meet in considering spontaneous radiation, for example. Suppose we start with the atom in an excited state, and with no energy in the radiation field. Then, after the transition, the atom will be in its normal state, having lost energy, and the radiation will be in an excited state, having gained the corresponding energy. The total energy of the system will be the same in either case. Now neither one of these situations is a steady state, for neither one persists indefinitely. Both are approximate steady states, corresponding to the same energy. The perturbation problem, then, is one in perturbations of a degenerate system, having two equal energy levels. We have seen that such a perturbation problem leads to mathematics just like two coupled mechanical systems, as two pendulums, and it is convenient to use the mechanical language in describing what happens. Our present problem is like two pendulums of equal period (corresponding to the equal energy levels), coupled together. If the first pendulum vibrates alone, that corresponds to the state in which the atom is excited; if the second vibrates, it corresponds to the radiation being excited. But neither of these mechanical motions can occur by itself; if we start one pendulum vibrating, in time it comes to rest, and the other takes up all the energy. This corresponds to the fact that the system gradually changes so that the atom is in its normal state, the radiation excited. There is a flaw in our analogy, however: the energy in the mechanical case goes back to the first pendulum, while the atom does not come back to the excited state. The answer to this difficulty is easily given. The radiation field actually has not one mode of motion only, but many,

all of about the same energy, all capable of interacting with the atom. Thus the emitted photon can travel in any direction, and not only that, photons of many different energies, all in the neighborhood of the energy ordinarily emitted by the atom, can interact, on account of the finite breadth of the spectral line. Thus while the situation where the atom is excited, and the radiation is in its normal state, is just one state, there are a great number of states corresponding to the other situation. It is as if our one pendulum corresponding to the excited atom, interacted with a great, or even infinite, number corresponding to the excited radiation. In these circumstances, the mechanical energy originally in the first pendulum would soon become dissipated, scattered through the others, and it will never happen all to come back to the first one, though a little might. Physically, the radiation emitted by the atom travels to a great distance, and is very unlikely ever to find its way back to the atom which sent it out. But if the whole thing is enclosed in a box with reflecting walls, there will be a certain chance, finite though small, that the radiation will be eventually reflected back to the atom and absorbed.

One significant feature of the situation is that there are real stationary states for the system of atom plus radiation. This follows directly from the fact that we can solve the perturbation problem. Just as with the coupled pendulums, there are normal coordinates, consisting of combinations of the various separate coordinates. Thus, there is some combination of all the various probabilities of the atom being in various states of excitation and the radiation field being in corresponding states which could persist indefinitely, and is thus a stationary state. The things we ordinarily think of as stationary states are combinations of these, just as the state where one pendulum is excited, the other at rest, is a combination of the two normal coordinates, with definite amplitudes and phases. These are really not stationary states at all, for they change with time. In any such problem, there are two equally good methods of treatment: first, we may use the unperturbed states which physically seem like stationary states, treating the perturbations between them by variation of constants, and so introducing apparent transitions into the problem; or secondly, we may introduce the real stationary states, by the ordinary perturbation theory, introducing the correct initial conditions, and following what happens as time goes on,

without having any transitions at all between these real stationary states. This point of view is very illuminating, for it shows us that the only distinction between stationary states and transitions is largely artificial, determined by the original unperturbed wave functions in which we choose to discuss the system.

239. Applications of Coupled Systems to Radioactivity and Electronic Collisions.—Many other problems of transitions can be looked at from the same point of view we have just used in discussing radiation. One example is the radioactive disintegration, which we have considered in Chap. XXIX, Fig. 64. We might take as approximate stationary states first the discrete states of a particle within the finite depression, second the continuous states of the particle outside. If the barriers were infinitely high, there would be no transitions between them, but if the barrier is finite, we may start with a particle within the nucleus, and consider that it has a certain probability of a transition to a state of equal energy outside the barrier. This could be treated by the perturbation theory of degenerate systems, where we could find the probability of leaking out by variation of constants, or alternatively could get approximations to the actual stationary states of the system. In this case, as with radiation, the probability of the particle coming back and getting back into the nucleus again, though small, is finite, if the system is enclosed in a finite box. Here the stationary states which are combinations of solutions for the discrete and continuous regions are perfectly reasonable and natural, and the more accurate way of solving the problem would be to determine these stationary states by the Wentzel-Kramers-Brillouin method, and build up a wave packet at $t = 0$ corresponding to having all the distribution inside the nucleus, and asking how this packet spreads out as time goes on, though without change of real stationary states.

Another similar problem is that of collisions, either elastic or inelastic. Suppose that an electron collides with an atom, being scattered either without change of energy, or with decrease or increase of energy corresponding to raising or lowering the energy of the atom. We can start with a number of unperturbed, not quite stationary, states: first, the electron approaching the atom, with the atom in its original state; secondly, an electron being scattered, say in a definite direction, or better with some function of angle represented by a spherical harmonic, with its initial energy, the atom being unchanged; thirdly, an electron

scattered with a decrease of energy corresponding to a transition of the atom, with the atom in the correspondingly excited state; fourthly, an electron scattered with increase of energy, the atom being in a lower state, after what is called a collision of the second kind. All these states have the same energy, so that the perturbation problem between them, resulting from the fact that they are not solutions of the problem in the region where the electron is in the atom, is one of transition between systems of the same energy. Here, as before, it is often convenient to proceed by the method of variation of constants, and from this we get the probabilities of the various elastic and inelastic impacts. One thing is worth noting in all these problems: in the method of variation of constants, the quantity determining the probability of transition is the nondiagonal matrix component of the perturbing energy between the different approximately stationary states. Thus the calculation resolves itself into a computation of these matrix components, and transitions are likely for which the matrix components are large. In our radiation problem, the matrix components in question were those of the electrical energy, involving directly the matrix components of electric moment of the atom.

While the perturbation method can be used for discussing collisions, it is not very accurate, on account of the large perturbations which the colliding electron exerts on the atom during the instant of collision. Fortunately, at least in the case of elastic collisions, much better approximation methods are available. As we shall see later, an atom acts on an electron very much like a central field of force, and the problem of the scattering of an electron by a central field is merely the special case of the central field problem, discussed in the next chapter, which we meet if the electron is in a continuous rather than a quantized energy level. By analogy with the results which we shall obtain in Sec. 241, the wave function of an electron in a central field is a product of spherical harmonics of angle, times a certain function of r , and for an electron coming from infinity, this function of r is of the form shown in Fig. 62, satisfying a definite boundary condition at the center of the atom, but becoming sinusoidal for large values of r . By combining an infinite number of such solutions, all corresponding to the same energy, but with different functions of angles, it can be shown that we can make the resultant wave at large distances approach a plane wave, representing a stream of electrons traveling in a definite direction. But the

functions are such that, if the central field is not vanishingly small, it is not possible to build up exactly a plane wave. Instead, there are certain terms left over representing spherical waves traveling outward from the center of force, with amplitudes proportional to $1/r$, so that they are negligible compared to the plane waves at sufficiently large distances. These spherical waves represent the elastically scattered electrons.

Two particularly interesting features of the elastic scattering can be investigated by the method just described. First, one may find the total intensity in the scattered wave, which can be proved to be equal to the total intensity removed from the plane wave by its passage over the atom. This gives the probability that an electron will be scattered by the atom, and it proves to increase as the atomic number of the atom increases, and to depend in a complicated way on the speed of the electron. This dependence is so complicated that in some cases, called the Ramsauer effect, very slow electrons have abnormally small probabilities of being scattered, and practically pass through the atom without hindrance. The probability of scattering is often described by defining an effective cross section for the atom, a cross section such that if all electrons striking it were scattered, and all passing around it were not, the probability of scattering would agree with the observed value. Plainly the effective cross section depends on electron velocity and on the nature of the atom. The second interesting feature of elastic scattering is the angular distribution of the scattered electrons, determined by the relative probabilities of scattering with the various spherical harmonic functions of angle. This again can show a complicated dependence on electron velocity and atomic constitution.

Problems

1. Prove that if both unperturbed and perturbed functions, u_n^0 and u_n , are orthogonal and normalized, the transformation coefficients S_{mn} satisfy the orthogonality and normalization conditions.
2. Show that if we expand the correct wave functions in a series of functions which are not exactly orthogonal or normalized, the equations for the transformation coefficients S_{mn} are

$$\sum_m (H_{km} - E_n d_{km}) S_{mn} = 0,$$

where $d_{km} = \int u_k^0 u_m^0 dv$, which now is not diagonal and is not equal to δ_{km} .

3. Consider a degenerate system in which there are two unperturbed wave functions, having equal diagonal energies $H_{11} = H_{22}$, which are nor-

malized but not orthogonal to each other, so that $\int \bar{u}_1^0 u_2^0 dv = d_{12} \neq 0$. Show that the two energy levels are $\frac{H_{11} + H_{21}}{1 + d_{12}}, \frac{H_{11} - H_{21}}{1 - d_{12}}$.

4. Show that the two correct wave functions in Prob. 3 are $\frac{u_1^0 + u_2^0}{\sqrt{2(1+d)}}$, respectively. Prove them to be normalized and orthogonal.

5. Solve the problem of a system with two degenerate unperturbed levels of the same energy, by the method of variation of constants. Show that the equations for the time derivatives of the c 's can be solved by assuming an exponential or sinusoidal solution. Show that the final solution is a pulsation from one state to the other, the frequency of pulsation being H_{12}/\hbar .

6. Prove by perturbation theory that the energy levels of a linear oscillator are not affected by a constant external field, except in absolute value, all being shifted up or down together. Why should this be expected physically?

7. Find whether a rotator's energy is affected, to the first or higher orders of approximation, by a constant external field in the plane of the rotator.

8. Prove in Einstein's derivation of Planck's radiation law that $B_{12} = B_{21}$, by considering equilibrium in the limiting case of extremely high temperature, noting that in this limit the probability of forced transition is large compared with that of spontaneous transition, on account of the large density of radiation.

9. Prove directly from Schrödinger's equation that the sum $\sum_n \bar{c}_n c_n$ always remains constant.

10. For the problem of interaction of atoms and radiation, when the atom starts in the m th state, work out $\bar{c}_m c_m$ as a function of time, and show that this, added to the other $\bar{c}_n c_n$'s, gives a constant.

CHAPTER XXXIII

THE HYDROGEN ATOM AND THE CENTRAL FIELD

In the preceding chapters we have been discussing the general principles and methods of wave mechanics. We have seen that from wave mechanics one can derive ordinary Newtonian mechanics as a special case. But by far the most interesting mechanical problem which demands wave mechanics for its solution is the structure of atoms, molecules, and matter in general. We shall accordingly devote the remaining chapters of this book to the structure of matter. This is a problem which is doubly interesting; first, as a most important subject in itself, secondly, as the finest illustration of wave mechanics.

240. The Atom and Its Nucleus.—An atom consists of a nucleus, and a number of electrons. All electrons are alike, electrified particles of negative charge $-e = -4.774 \times 10^{-10}$ e.s.u., mass of 9.00×10^{-28} gm. Nuclei are heavier, and positively charged. The charges on nuclei are found in every case to be integral multiples of the charge e . Thus a nucleus may have a charge Ze , where Z is an integer, and in this case Z is called the atomic number. If the atom has enough electrons to be electrically neutral, it is obvious that it must have Z electrons, so that the atomic number measures both the charge on the nucleus and the number of electrons in the neutral atom. We shall see that this number Z is the determining quantity in fixing the properties of the atoms; if all atoms are tabulated in order of their atomic numbers, they show periodic properties, for reasons which we shall discuss in the next chapter, and this arrangement is called the periodic table of the elements. Of course, the number of electrons on the atom does not always have to be just the atomic number; violent methods, as bombardment, can knock electrons off, or in some cases extra electrons can be added, producing positive or negative ions, respectively. We shall see that some elements, the electropositive or alkaline ones, have a tendency to lose electrons, and form positive ions, while the basic elements tend to gain electrons and become nega-

tive ions. Atoms often enter chemical compounds as ions, rather than neutral atoms, so that in our study of atomic structure we shall have to speak constantly of ions as well as neutral atoms.

The element of atomic number one is hydrogen, the simplest element. Its nucleus is an elementary particle, called the proton, with mass 1,846 times that of the electron. The heavier nuclei appear to be built up from a combination of protons and neutrons, particles of no charge, but of mass approximately equal to that of the proton. There are approximately equal numbers of protons and neutrons in any nucleus, making the atomic weight (the mass of the nucleus, in multiples of the mass of the proton) approximately twice the atomic number, though this rule is far from exact, the heavier atoms containing more neutrons in proportion than the light ones. The forces holding the nucleus together are presumably largely forces of attraction between protons and neutrons, more than counterbalancing the repulsions between protons on account of their like electric charges. By the action of these forces, stable structures are produced, disintegrating only in the case of the heavy, radioactive elements, or in the very light elements under heavy bombardment. The theory of the structure of the nucleus is still in a preliminary state, and we shall not consider it; ordinary properties of matter prove to be almost completely independent of the nuclear structure, depending only on its charge and mass, with most properties depending only on its charge, so that two nuclei of the same charge and different masses, called isotopes, exhibit almost identical properties. Such isotopes are of very common occurrence, many ordinary elements being a mixture of several, the chemical atomic weights being weighted means of the weights of the isotopes, explaining why many observed atomic weights are far from whole numbers.

241. The Structure of Hydrogen.—The simplest element is hydrogen, with but one electron moving about a single nucleus. Fortunately the problem of its structure, according to wave mechanics, can be exactly solved, and it serves as a model for the more complicated elements. In fact, we have already carried out many of the mathematical steps in problems at one time or another, so that we shall merely have to summarize results here. For generality, we shall treat not merely hydrogen, but the problem of a single electron moving about a nucleus of charge Ze . The first thing we notice is that the nucleus is very heavy, com-

pared with an electron. Now if we have a single electron and a single nucleus, exerting forces on each other, we find, in wave mechanics as in classical mechanics, that the center of gravity of the system remains fixed, each particle moving about the common center of gravity. But the center of gravity is very close to the nucleus; it divides the vector joining nucleus and electron in the ratio of 1:1,846. Thus the nucleus executes only very slight motions, and practically we can treat it as being fixed, and the electron as moving about a fixed center of attraction. We shall find that this is a very general method in discussing the structure of matter: we first assume all nuclei to be fixed, and discuss the motion of the electrons about them. Only later do we have to take the motions of the nuclei into account. We discuss this more in detail in a later chapter.

We have, then, an electron of charge e , mass m , moving in a central field of force. The attractive force of the nucleus has a potential energy $-(Ze^2/r)$. Thus Schrödinger's equation, with the time eliminated, is

$$Hu = \left(-\frac{\hbar^2}{8\pi^2m} \nabla^2 - \frac{Ze^2}{r} \right) u = Eu. \quad (1)$$

We shall find it convenient in all our atomic problems to introduce at the outset so-called atomic units of distance and energy. The unit of distance is $a_0 = \hbar^2/4\pi^2me^2$, a unit first introduced in Bohr's theory of the hydrogen atom, but which comes into the present discussion as well. It is equal to 0.53 Ångström. The unit of energy most convenient to use is $2\pi^2me^4/\hbar^2$, though sometimes a unit twice as great is used. This is the energy required to ionize a hydrogen atom from its normal state. It is most conveniently stated, not in ergs, but in volt-electrons. A volt-electron by definition is the energy an electron acquires in falling through a difference of potential of 1 volt, or $eV = 4.774 \times 10^{-10} \times \frac{1}{3.10} \text{ ergs}$. In terms of this, our fundamental unit of energy is 13.54 volt-electrons. Associated with this energy is a frequency, given by energy $= h\nu$, and a wave length, and its reciprocal a wave number, given by $1/\lambda = \nu/c$. The wave number associated with our unit of energy is the so-called Rydberg number, $R = 109,737 \text{ per centimeter}$, and the corresponding energy is Rhc .

In terms of our atomic units, Schrödinger's equation for hydrogen can be rewritten, eliminating all the dimensional constants.

Thus, if our new distances are the old ones divided by a_0 , the new energy the old divided by Rhc , we easily find that

$$\left(-\nabla^2 - \frac{2Z}{r}\right)u = Eu, \quad (2)$$

where the derivatives are to be taken with respect to the new x, y, z . The coefficient 2 in the potential energy appears in the process of changing variables, the potential energy of two electronic charges being $2/r$ in these units.

Schrödinger's equation can now be solved, in spherical coordinates, by separation of variables. Using the results of Chap. XV, Probs. 6 to 8, the equation can be separated, letting $u = R\Theta\Phi$, and the differential equations are

$$\begin{aligned} \frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) + \left[E + \frac{2Z}{r} - \frac{l(l+1)}{r^2} \right] R &= 0, \\ \frac{1}{\sin \theta} \frac{d}{d\theta} \left(\sin \theta \frac{d\Theta}{d\theta} \right) + \left[l(l+1) - \frac{m^2}{\sin^2 \theta} \right] \Theta &= 0, \\ \frac{d^2\Phi}{d\phi^2} + m^2\Phi &= 0. \end{aligned} \quad (3)$$

The solutions of the second and third are $\Theta = P_l^m(\cos \theta)$, $\Phi = e^{\pm im\phi}$, or $\cos m\phi$ or $\sin m\phi$, where m must be an integer in order to have the function single-valued as far as ϕ is concerned, and l must be an integer in order not to have the function P become infinite for $\cos \theta = 1$. The P 's are called associated spherical harmonics, and are given by

$$\begin{aligned} P_l^m(\cos \theta) &= \sin^{|m|} \theta (A_0 + A_1 \cos \theta + A_2 \cos^2 \theta + \dots), \\ A_k &= A_{k-2} \frac{(k + |m| - 1)(k + |m| - 2) - l(l+1)}{k(k-1)}. \end{aligned} \quad (4)$$

For integral l 's, this series breaks off, the last nonvanishing term being for $k = l - |m|$. For even $l - |m|$, the expansion is in even powers, and for odd $l - |m|$ in odd powers. The functions R are discussed in Prob. 3. We use a simple transformation of the dependent variable, $y = rR$. The equation in this variable is

$$\frac{d^2y}{dr^2} + \left[E + \frac{2Z}{r} - \frac{l(l+1)}{r^2} \right] y = 0. \quad (5)$$

The solution is

$$\begin{aligned} y &= e^{-r\sqrt{-E}} r^{l+1} (A_0 + A_1 r + A_2 r^2 + \dots), \\ A_k &= -2A_{k-1} \frac{Z - (l+k)\sqrt{-E}}{(l+k)(l+k+1) - l(l+1)}. \end{aligned} \quad (6)$$

This series breaks off if $E = -Z^2/n^2$, where n is an integer. A simple discussion shows that if it does not break off, the resulting infinite series becomes infinite as r becomes infinite like $e^{2r\sqrt{-E}}$, so that y becomes infinite, and is not admissible as a wave function for a stationary state. We therefore limit ourselves to integral n 's, and n is called the principal or total quantum number, determining the energy. In terms of it, we have

$$y = e^{-\frac{rZ}{n}} r^{l+1} (A_0 + A_1 r + \cdots + A_{n-l-1} r^{n-l-1}),$$

$$A_k = -\frac{2Z}{n} A_{k-1} \frac{n-l-k}{(l+k)(l+k+1)-l(l+1)}. \quad (7)$$

From this recursion formula, we see that l cannot be greater than $n-1$, in order to have any terms to the series; and from the earlier recursion formula for the function of θ , $|m|$ cannot be greater than l . The principal quantum number n , and the so-called azimuthal quantum number l must both be positive, the smallest allowable value of n being 1 and of l zero. The so-called magnetic quantum number m , however, can be positive, negative, or zero, so long as its magnitude falls within the allowed limits.

242. Discussion of the Function of r for Hydrogen.—Though we have an exact solution for hydrogen, a qualitative discussion is still desirable, using the method of energy. In Chap. VII we have already discussed motion in a central field in classical mechanics. We have seen that the motion along the radius is like a one-dimensional oscillation, in a potential field $V + p^2/2mr^2$, where V is the potential energy, p the angular momentum. In our case, the differential equation for y is like a one-dimensional wave mechanical problem with a potential, in atomic units, of $-\frac{2Z}{r} + \frac{l(l+1)}{r^2}$, or in ordinary units $-\frac{Ze^2}{r} + \frac{p^2}{2mr^2}$, where $p = \sqrt{l(l+1)} \frac{h}{2\pi}$. It is thus clear in the first place that the quantum number l determines the angular momentum, in units of $h/2\pi$, though the values are not l times this unit, but $\sqrt{l(l+1)}$ times it. We shall further discuss the angular momentum later on. Now it is interesting to draw the various potentials, as we do in Fig. 66, where $-\frac{2}{r} + \frac{l(l+1)}{r^2}$ is plotted, for $l = 0, 1, 2$. We have also plotted $-\frac{2}{r} + \frac{k^2}{r^2}$,

indicated by the dotted lines. The reason for this is that in Bohr's theory of hydrogen, it was assumed that the electron moved according to classical mechanics, and that its energy could have only those particular values for which the quantum conditions were fulfilled. He assumed that the angular momentum was $kh/2\pi$, where k was an integer, so that if we discuss

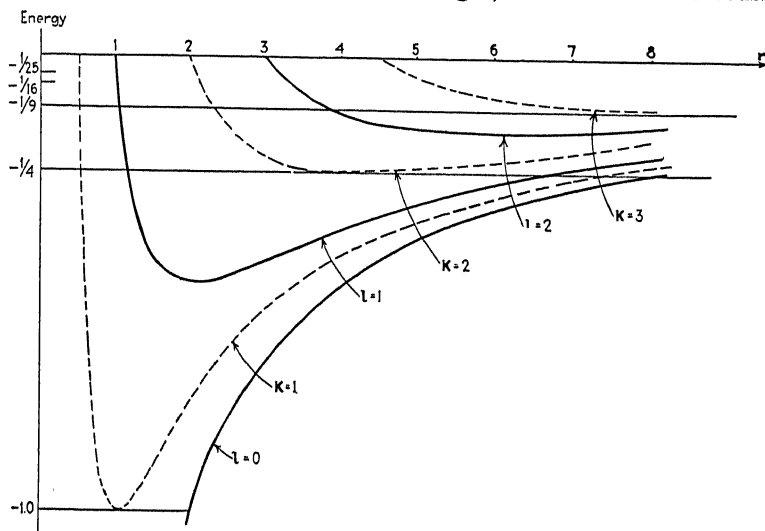


FIG. 66.—Potential and energy levels for hydrogen. Full lines: $-\frac{2}{r} + \frac{l(l+1)}{r^2}$ (potential corrected for centrifugal force, wave mechanics). Dotted lines: $-\frac{2}{r} + \frac{k^2}{r^2}$ (corrected potential, Bohr theory). Horizontal lines represent energy levels.

the classical motion with these dotted potential curves, we shall have precisely Bohr's orbits. He also assumed

$$\oint p_r dr = \oint \sqrt{2m(E - V - p^2/2mr^2)} dr = n_r h,$$

where

$$p = kh/2\pi.$$

The energy levels, either on Bohr's theory or wave mechanics, are $-1/n^2$, where on Bohr's theory $n = k + n_r$, and these are drawn, at -1 , $-1/4$, $-1/9$, etc. Now consider the particular case $k = 1$. The lowest possible energy level for this is evidently -1 ; for here E intersects the potential curve at but one point, giving, therefore, a circular orbit, the perihelion and aphelion distances being equal. As we see from the diagram, the radius

of the circular orbit is one unit, and the energy minus one unit, explaining, therefore, the origin of the units. But for this same k , higher energy levels are connected with elliptical orbits, as, for example, that for which $n = 2$, $k = 1$, with perihelion smaller, aphelion larger than the circle for $n = 1$. For $n = 2$ there is a second Bohr orbit, for $k = 2$: a circle of radius 4 units. Similarly for $n = 3$, there are three orbits, for $k = 1, 2, 3$, and so on, the orbit for $k = n$ being in each case a circle. This question is discussed in a problem, where it is shown that the orbits are ellipses, of semimajor axis equal to $\frac{n^2}{Z} \frac{h^2}{4\pi^2 m e^2} = \frac{n^2}{Z} a_0$, and minor axis equal to k/n times the major axis.

In the wave mechanics, where the angular momentum has the nonintegral value $\sqrt{l(l+1)}$ units, we must use the full lines. Now we are interested in the region where the kinetic energy is positive, not as the only place where motion can occur, but as the region where the wave function is sinusoidal. Outside this region, it falls off exponentially. We can see a few examples in Fig. 67, in which the first few wave functions are plotted (we plot y , equal to r times the radial part of the wave function). On each function the limits of the region of classical motion are determined by the fact that the points of inflection come here, the tendency of the curves being sinusoidal between the points, exponential outside. It is plain that the wave functions are larger where the electron is likely to be found, small where it is not, as we could prove by deriving the solution from the Wentzel-Kramers-Brillouin method, a possible, though not very convenient, method of discussing the hydrogen problem. As this method would show at once, the wave length and amplitude both become large as r becomes large, and $E - V$ becomes small, so that the outermost maximum of the wave function is in all cases the largest, and contributes most to the wave function as a whole. One property of the wave function is evident from Fig. 67: for small r , the behavior is determined mostly by l , for large r mostly by n . This is natural from the fact that for small r the quantity $E + \frac{2}{r} - \frac{l(l+1)}{r^2}$ approaches $-\frac{l(l+1)}{r^2}$, and for large r it approaches $E + \frac{2}{r} = -\frac{1}{n^2} + \frac{2}{r}$.

We note that as l becomes smaller and smaller, the region where the wave function is large, or the classical orbit, penetrates

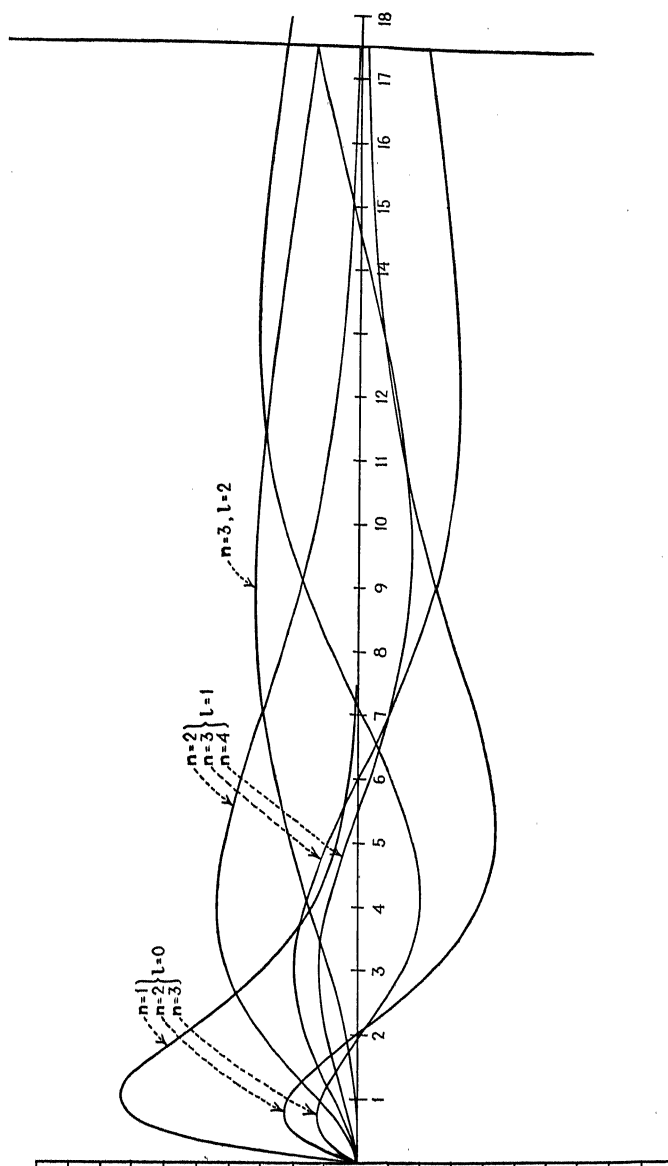


Fig. 67.—Radial wave functions for hydrogen. Functions plotted are r times the radial part of the wave function.

closer and closer to the nucleus. For large r , and, as a matter of fact, for the whole outer maximum, which, as we have seen, is the most important one, a fairly good approximation to the wave function is simply $r^n e^{-\frac{Zr}{n}}$, the wave function for the orbit of maximum azimuthal quantum number ($l = n - 1$), corresponding to the circular orbit in Bohr's theory. It is interesting to note that this function has its maximum at $r = \frac{n^2}{Z}a_0$, just the radius of the corresponding circular orbit in Bohr's theory.

243. The Angular Momentum.—We have seen that the quantity $\sqrt{l(l+1)}\frac{h}{2\pi}$ corresponds to the angular momentum of the orbit. This can be seen by computing the matrix of total angular momentum, or rather of its square, which is more convenient. We can most easily get the operator for the angular momentum, in spherical coordinates, by an indirect method. Classically, $H = p_r^2/2m + p^2/2mr^2 + V$, where p is the total angular momentum. Now in wave mechanics we find the wave equation such that

$$H = \frac{1}{2mr^2} \frac{h}{2\pi i} \frac{\partial}{\partial r} \left(r^2 \frac{h}{2\pi i} \frac{\partial}{\partial r} \right) + \frac{1}{2mr^2} \left[\frac{1}{\sin \theta} \frac{h}{2\pi i} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{h}{2\pi i} \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \left(\frac{h}{2\pi i} \right)^2 \frac{\partial^2}{\partial \phi^2} \right] + V.$$

By comparison, it is plain that the operator for p^2 is

$$p^2 = \left(\frac{h}{2\pi i} \right)^2 \left[\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right]. \quad (8)$$

But now from the differential equations for Θ and Φ , we easily have, using this operator,

$$p^2 u = l(l+1) \left(\frac{h}{2\pi} \right)^2 u. \quad (9)$$

That is, p^2 has a diagonal matrix (since $p^2 u$ is a constant times u , without any terms in other characteristic functions), and the diagonal value is $l(l+1)(h/2\pi)^2$, so that the total angular momentum is constant, as it must be in the absence of torques. We can also easily find the component of angular momentum along the z axis. The angular momentum along this axis is the momentum conjugate to the angle ϕ of rotation about the axis, so that its operator is $\frac{h}{2\pi i} \frac{\partial}{\partial \phi}$. Now take the solutions where ϕ enters

into the wave function as the exponential, $e^{\pm im\phi}$. Then $p_\phi u = \frac{h}{2\pi i} \frac{\partial u}{\partial \phi} = \pm m \frac{h}{2\pi} u$. This again is diagonal, showing that the component of angular momentum remains constant. Further, if we use the wave function $e^{im\phi}$, the component equals $m h/2\pi$.

The interpretation of these results is best made in terms of a vector model. Suppose we consider that the angular momentum of the orbit is $l h/2\pi$. This will then be regarded as a vector, normal to the plane of the orbit, pointing in some arbitrary direction in space. The component of angular momentum along the z axis is simply the projection of the vector in that direction. Now we find that this can have only the quantized values $m h/2\pi$. Hence there are only a finite number of possible orientations for the orbit, as shown in Fig. 68, for the states for $l = 3$. Plainly m can go from a maximum of l to a minimum of $-l$, or $2l + 1$ values in all, just as one finds from the discussion of the spherical harmonics. Now this vector diagram is only suggestive, not strictly true. We see this from the fact that our vector has

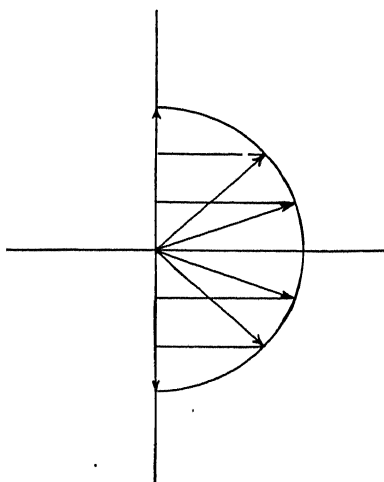


Fig. 68.—Possible orientations of angular momentum, for $l = 3$.

length $l h/2\pi$, while the actual angular momentum is $\sqrt{l(l+1)} h/2\pi$. The fundamental reason is that, since the angular momentum and its component are exactly given, the uncertainty principle does not allow us to fix definitely the plane of the orbit, which corresponds to a coordinate. As a matter of fact, the electron in wave mechanics does not move exactly in a plane, but strays outside the plane, as the uncertainty principle would suggest. This is best shown by polar diagrams of the spherical harmonics, plotting the square of the spherical harmonic, which gives the density, as function of angle. This is done in Fig. 69, for $l = 1$, $m = 1$ and 0 , and $l = 2$, $m = 2, 1, 0$. ($l = 0$ does not depend on angle.) If we imagine these figures rotated about the axes, we see that for $m = l$, the figure indicates that most of the

density is in the plane normal to the axis, but considerable is out of the plane. For $l = 2$, $m = 1$, for instance, the density lies near a cone, as if the plane of the orbit took up all directions whose normal made the proper angle with the axis.

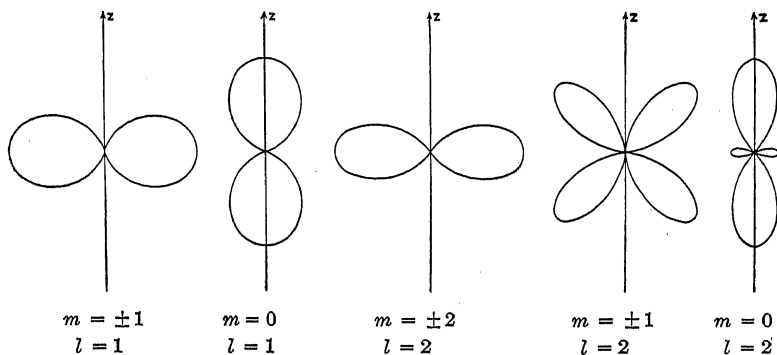


FIG. 69.—Dependence of wave functions on angle. Θ^2 plotted in polar diagram.

244. Series and Selection Principles.—All the states for a given value of l and n , but different m , have the same function of r , and the same energy. We shall find that this is still true with an arbitrary central field, so that even in that problem the solution is degenerate. Physically, so long as the angular momentum is determined, it cannot make any difference as far as the energy is concerned which way the orbit is orientated, on account of the spherical symmetry. Thus we often group together the various substates with the same l and n but different m , regarding them as constituting a single degenerate state, with a $(2l + 1)$ fold degeneracy. For hydrogen, the energy as a matter of fact depends only on n , so that all states of the same n but different l values are degenerate, but this is not true in general for a central field. It is convenient, rather, to group all the states of the same l value but different n together to form a series, since they are closely connected physically, having the same functions of angle, while those of the same n merely happen to have the same energy, but without important physical resemblances. The series of different l values are conventionally denoted by letters, derived from spectroscopy. We have the table as shown on page 417. By order of degeneracy we mean simply the number of sub-levels of different m values.

The classification into series becomes important when we consider the transition probabilities from one level to another. We

l value	Letter	States	Order of degeneracy
0	s	$1s, 2s, 3s, \dots$	1
1	p	$2p, 3p, 4p, \dots$	3
2	d	$3d, 4d, \dots$	5
3	f	$4f, 5f, \dots$	7
4	g	$5g, \dots$	9

recall that these are given by the matrix components of the electric moment between the states in question. When these components are computed, it is found that there are certain selection rules:

1. The component is zero unless the l 's of the two states differ by ± 1 unit.

2. The component is zero unless the m 's differ by 0 or ± 1 unit.

The latter rule is easily proved. For, suppose we compute the matrix components of $x + iy$, $x - iy$, z , which are simple combinations of x , y , z , the three components of displacement. If we find the matrix components of all three of these to be zero for a given transition, the transition will be forbidden. Now these three quantities, in polar coordinates, are $r \sin \theta e^{i\phi}$, $r \sin \theta e^{-i\phi}$, $r \cos \theta$, respectively. If u is $R\Theta e^{im\phi}$, we have $(x + iy)u = rR \sin \theta \Theta e^{i(m+1)\phi}$, showing that this quantity has a matrix component only to states having the quantum number $m + 1$, since the quantity on the right could be expanded in series of functions with many values of n and l , but only the one value $m + 1$. Similarly $(x - iy)u = rR \sin \theta \Theta e^{i(m-1)\phi}$, allowing transitions only from m to $m - 1$, and $zu = rR \cos \theta \Theta e^{im\phi}$, allowing only transitions in which m does not change. The proof of the selection principle for l is slightly more difficult, involving the theorem that $\sin \theta P_l^m(\cos \theta)$ or $\cos \theta P_l^m(\cos \theta)$ can be expanded in spherical harmonics whose lower index is $l + 1$ or $l - 1$ only.

The selection rules have the following results: If we arrange the series in order, *spdf* . . . , a level of one series can only have transitions to the immediately adjacent series. This gives us the transitions indicated in Fig. 70 (all of the transitions between upper states are not indicated; merely some of the more important ones down to lower states). The series of lines arising from transitions of the p states to $1s$ is called the principal series; from the s terms to $2p$, the sharp series; from the d terms to $2p$, the

diffuse series; from the f terms to $3d$, the fundamental series. The letters s , p , d , f are the initials of these series. When the matrix components are worked out, the strongest lines are those in which l decreases by one unit (principal, diffuse, and fundamental series), and those for which l increases (as the sharp series) are weaker. Of course, on account of the degeneracy in l in hydrogen, the different series are not separated, but they are in other atoms, and it is for those that the classification is impor-

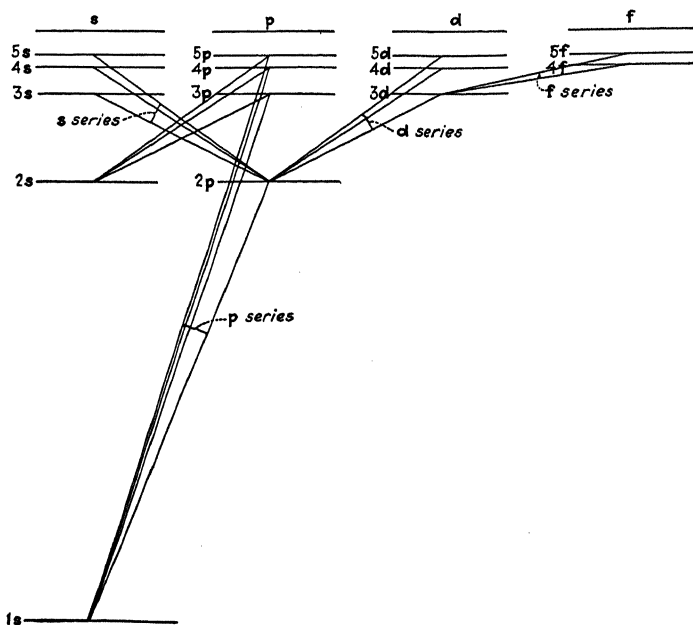


Fig. 70.—Energy levels and allowed transitions and series in hydrogen.

tant. To see this, we must study the energy levels in the general central field.

245. The General Central Field.—We shall find that in discussing atomic structure, we shall wish to consider that each electron moves in a central field, but not an inverse square field. The field is rather the sort which we should have if there were a nucleus of charge Z units, surrounded by a spherical ball of negative charge, having a total charge $-(Z - 1)$ units, corresponding to the remaining electrons of the atom. Such a field has a potential $\frac{2Z(r)}{r}$, where $Z(r)$ goes from 1 at large r to Z at

small r . For such a potential, most of our discussion goes through without alteration. The differential equation can be separated in the same way, and the functions of angle are just the same, so that our classification into series, vector model, and selection principles holds as with hydrogen. The only difference comes in the function of r , and in the values of the energy levels. We can no longer solve the equation exactly, and shall use the qualitative method of discussion. In Fig. 71 we show a diagram, like Fig.

66, in which we plot $-\frac{2Z(r)}{r} + \frac{l(l+1)}{r^2}$. The potential is so

chosen that for r greater than unity, $Z(r)$ is just unity, but for smaller r 's $Z(r) = 10 - 9r$, so that the charge approaches 10 at $r = 0$, but joins on smoothly at $r = 1$. It is obvious that the s electrons are greatly affected by the change in potential. The $1s$ wave function is located practically all inside $r = 1$. Thus its potential curve is practically $\frac{-2(10 - 9r)}{r} = -\frac{2(10)}{r} + 2(9)$

for the whole range. In other words, it is like a hydrogen problem of nuclear charge 10 units, but with the constant correction $2(9)$ to be added to the energy. The energy of such a state would be $-(10)^2 = -100$, and when we add our constant 18, it is -82 units, showing that this level is very tightly bound. Similarly the $2s$ is largely inside, though not so completely, and to a somewhat poorer approximation its energy is $-\frac{100}{4} + 18 = -7$ units.

The higher s orbits, however, project out into the region beyond $r = 1$, where the potential is hydrogen-like with charge 1, and we shall discuss them in a moment. The p , d , . . . states, on the other hand, are almost entirely outside the range where the potential is not hydrogen-like. Their energy levels and wave functions are almost exactly like those of hydrogen.

It is seen from this discussion that we can divide the levels in such a case into three classes: (1) those entirely inside the range of large potential, which will prove to be those inside the atom; (2) those half in and half out; and (3) those entirely outside. The levels of larger l values do not penetrate the inside, and belong to group 3. In this case, we reach this situation with $l = 1$, but with larger cores of negative charge about the nucleus, and so larger regions where the potential is much greater numerically than in hydrogen, the p electrons, or in some cases the d or even f electrons, are penetrating. For the lowest l values, in any

case, the orbits of large n are partly outside but penetrate inside, and those of small n are entirely inside the core of negative charge. These penetrating orbits have quite different energy values from the nonpenetrating ones, so that the different series do not lie

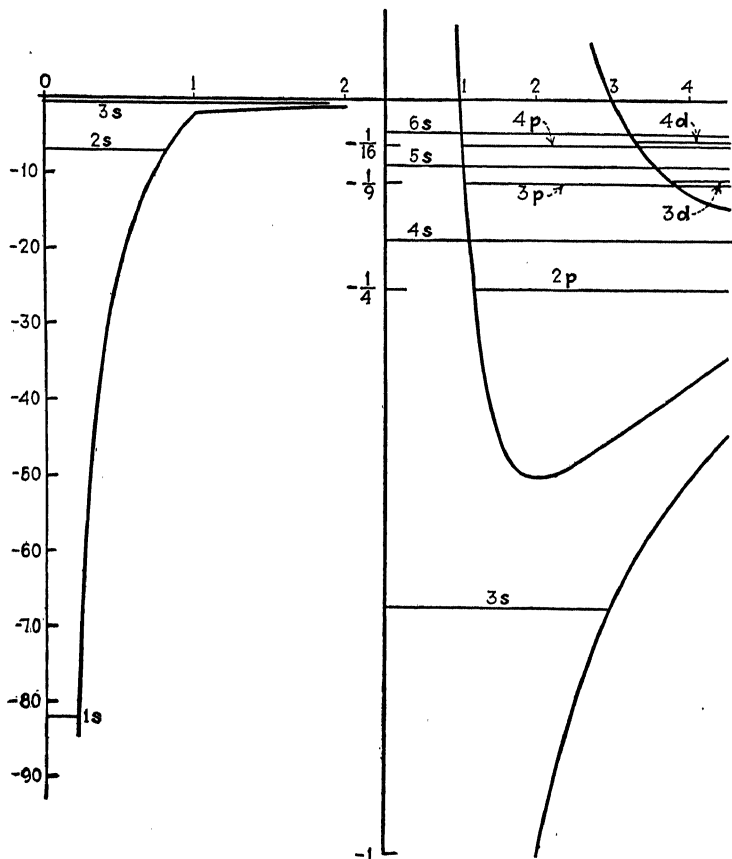


FIG. 71.—Potential and energy levels for a central field, with $Z(r) = 10 - 9r$ from $r = 0$ to 1 , $Z(r) = 1$ for r greater than unity. Left-hand diagram on different energy scale.

on top of each other, as in hydrogen. For the orbits which penetrate in their inner parts only, we get a formula for the energy, from the quantum condition. This formula is most conveniently derived using Bohr's form of the azimuthal quantum condition. We have $\int p_r dr = n_r h$ for the radial quantum condition. Then for hydrogen, $\left(\int p_r dr + kh \right) = nh = \frac{1}{\sqrt{-E}} h$,

where k is Bohr's azimuthal quantum number. Thus $\int p_r dr = \frac{h}{\sqrt{-E}} - kh$. For a penetrating orbit with our form of potential, the integral over the outer part of the orbit, where the potential, and hence p_r , are hydrogen-like, will have just the same value as here, if we use the proper energy. For the inside, however, p_r is much greater, so that there is an additional contribution to the integral, as we see from Fig. 72. This contribution, moreover, is roughly the same for all terms of the same k value, since the

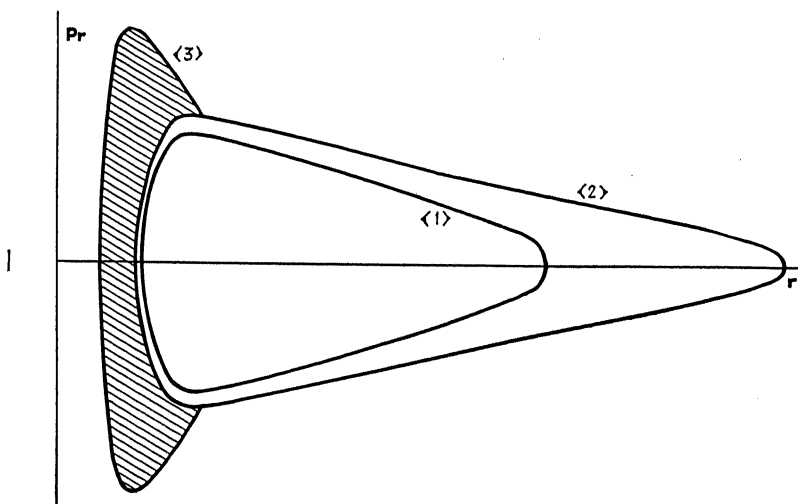


FIG. 72.—Phase space and phase integral for r , penetrating and nonpenetrating orbits. (1) and (2): Nonpenetrating orbits of same k , different n . (3) combined with (2): Penetrating orbit, having same energy as (2), but in a non-Coulomb field, so that it has a different quantum number and phase integral. Shaded area represents the quantum defect δ .

inner part of the orbit depends almost entirely on the angular momentum alone. Thus we have for the general case

$$\int p_r dr = \frac{h}{\sqrt{-E}} - kh + \delta_1 h,$$

where δ_1 is a function of k only, to the first approximation. The result must be $n_r h$, by the radial quantum condition, so that we have

$$E = -\frac{1}{(n_r + k - \delta_1)^2} = -\frac{1}{(n - \delta_1)^2}, \quad (10)$$

where n is the total quantum number, and where δ is called the quantum defect. A more careful discussion, using the Wentzel-Kramers-Brillouin method, shows that the same formula still holds when we use $\sqrt{l(l+1)}$ in place of k , and remember that we must use half quantum numbers. This formula, which can be written, in wave numbers, $E = -\frac{R}{(n - \delta_1)^2}$, is called Rydberg's formula, and was first discovered experimentally by Rydberg. We see then that the penetrating orbits fall into series as the nonpenetrating ones do, but that we must subtract the quantum defect from the quantum numbers. These quantum defects range from 0 for the nonpenetrating orbits to sometimes quite large values, even of the order of 5 or 6, for the s electrons of heavy atoms. From experimental observations of spectral series, we can find the quantum defects, and so tell which orbits are penetrating, and which are not. In the next chapter we shall discuss in more detail the energy levels for the orbits entirely inside the atom, which are most directly concerned in atomic structure.

The wave functions for the central field of the type we are discussing are not very different in general from those for hydrogen. But there are important differences in detail. We note that a hydrogen-like orbit corresponding to the problem of nuclear charge Z is $1/Z$ times as great as that for nuclear charge 1. Hence, in the case of Fig. 71, the $1s$ and $2s$ orbits are something like $1/10$ as large as for hydrogen. The penetrating orbits, like $3s$, $4s$, etc., will have the inner loops small in proportion, as the $1s$ and $2s$ are, but the outer parts, being in a field of charge 1, will be large. Thus there will be a much greater disparity between the size of the inner and outer loops than even for hydrogen, the outer ones being much more important in consequence. We may see this from the Wentzel-Kramers-Brillouin method. Here both amplitude and wave length go inversely with p_r . In the penetrating part of the orbit, p_r is much greater than for hydrogen, for the same total energy, so that amplitude and wave length become extremely small. The physical way to say this is that the electron moves very fast when it penetrates the core and is exposed to the whole charge of the nucleus, and hence spends but a very short time there, so that the wave function is small. For actually computing the wave functions, we can best use numerical integration of the differential equation, or the

method of Wentzel, Kramers, and Brillouin. We shall discuss wave functions more in detail in the next chapter.

Problems

1. Work out the spherical harmonics for $l = 3$, and draw diagrams for them similar to Fig. 69.

2. Prove from the differential equation that the associated spherical harmonics are orthogonal. Verify this for the cases of $l = 1$ and 2.

3. Carry out the solution of the radial wave function for hydrogen, deriving Eqs. (5), (6), and (7), following the method outlined in the text, and verifying that if the series does not break off it diverges.

4. Show that $y^2 dr$, where $y = rR$, is proportional to the probability of finding the electron between r and $r + dr$. Compute radial wave functions for states $1s$, $2s$, $3s$ for hydrogen, and draw graphs of y^2 .

5. Prove that for a radial wave function without nodes ($l = n - 1$), for nuclear charge Z , the maximum of y comes at n^2/Z .

6. Using the results of Prob. 3, Chap. IX, set up the radial phase integral for Bohr's model of hydrogen, showing that $E = -1/n^2$. Using the properties of the ellipse mentioned in Prob. 4, Chap. VII, verify the statements of Sec. 242 regarding the dimensions of the orbits.

7. Draw an energy level diagram in which the substates of different m 's are shown, drawing them as if slightly separated, including states $1s$, $2s$, $3s$, $2p$, $3p$, $3d$. Indicate all transitions allowed by the selection principles for l and m , as in Fig. 70.

8. Prove that the potential used in Fig. 71 is what would be found with a nucleus of 10 units charge, surrounded at distance unity by a hollow sphere, with 9 units of negative charge uniformly distributed over the surface.

9. A rough model of the inner electrons of the sodium atom can be obtained by assuming the nucleus of charge 11 units; a shell of radius 0.09 units, with two electronic charges spread over the surface; and a shell of radius 0.58 units, with 8 electrons spread over it, so that the net charge is 1 unit positive. Set up a diagram like Fig. 71 for such a potential field, drawing the potential functions for s , p , d electrons. Find which orbits are nonpenetrating.

10. Using the potential of Fig. 71, and Bohr's azimuthal quantum condition, compute the positions of $3s$, $4s$, and $5s$ levels. To do this, evaluate the radial quantum integral, computing separately the parts inside and outside $r = 1$, set the sum equal to $n_r h$, and solve for the energy, using numerical methods if necessary to solve the transcendental equation. Find how closely the result fits with the Rydberg formula, computing quantum defects for each level.

11. In the field of Fig. 71, the p electrons do not have exactly the hydrogen energies, for their wave function is not zero in the region inside $r = 1$, where the potential is not hydrogen-like. Compute the first-order perturbed value of the energies of $2p$, $3p$, $4p$, by using hydrogen wave functions as the starting point of a perturbation calculation, and assuming the difference between the hydrogen potential and the actual one as perturbative potential.

Compute quantum defects for each level, seeing how well the Rydberg formula is obeyed. It is to be noted that in such a case as this, the second-order perturbation is often more important than the first, so that our calculation is not very accurate.

12. Apply the Wentzel-Kramers-Brillouin method to the wave functions of hydrogen, computing approximate radial functions for $3p$, $4p$, and comparing with the exact solutions.

CHAPTER XXXIV

ATOMIC STRUCTURE

The electrons in an atom move, to an approximation, in central fields of force, each in the field produced by the nucleus and the average charge of the other electrons. Thus, as we have seen in the last chapter, there are different quantum numbers which they can have. We can have in an atom $1s$, $2s$, $2p$, . . . electrons. All electrons of a given total quantum number, inside the atom, have roughly the same radius for the maximum of their wave functions, and roughly the same energy, in contrast to the electrons which are largely outside, in which s and p electrons are more tightly bound on account of penetration. We can then group the electrons of the same total quantum number together into shells, those of $n = 1$ forming what is called the K shell, those with $n = 2$ the L shell, $n = 3$ the M shell, etc., the letters K , L , M , . . . coming from x-ray notation. The inner electrons are the most tightly bound and hardest to remove, and hence connected with the highest frequencies in the spectrum: the K series of x-rays, connected with the electrons of the K shell, has shortest wave length, L series next, and so on. On the other hand, an outer electron is shielded from the nuclear attraction by the presence of the other electrons; for the electrical force acting on a charge in a spherical distribution is what we should have if we imagined a sphere drawn about the center through the charge in question, forgot about all charge outside this sphere, imagined the charge inside the sphere concentrated at the center, and calculated its attraction by the inverse square law. Thus for an inner electron we forget about almost all the other electrons and have practically the unadulterated attraction of the nucleus, but with an outer electron the number of other electrons within the sphere is almost equal to the nuclear charge and almost cancels it, leaving only a small net attraction, and an easily detached electron. It is convenient in this connection to speak of an "effective nuclear charge" Z_e , and a shielding constant S ; Z_e is the charge which, placed at the center, would produce the

same attraction as the nucleus and electrons, and thus varies from Z for the inner electrons down to the order of magnitude of 1 for the outer ones, and S is defined by $Z_e = Z - S$, so that S measures roughly the number of electrons inside the sphere in question. In general, we see that each electron in an atom, or at least each shell, will have a different shielding constant. And now it is an important fact that the energies involved in ordinary chemical and physical processes are only large enough to remove or disturb the outer electrons of an atom, and leave the inner ones unaffected. Only x-rays, very violent bombardment, and such extreme means can disturb the inner electrons, and as a result we need not consider them in ordinary chemical and physical applications.

246. The Periodic Table.—The series K, L, M, \dots of shells has no obvious end, and yet an atom has but a finite number of electrons. It is evident, then, that the shells cannot all be filled. The attraction of the nucleus will pull electrons into the lowest shells, until they are filled, and then the rest will have to go into higher ones. The capacity of a shell is strictly limited, according to a very important principle called the exclusion principle (excluding more than a certain number of electrons from a shell), so that a K shell can contain only 2 electrons, an L shell 8, an M shell 18, an N shell 32, and so on. Using this principle, we can begin to see how the atoms build up, and in so doing we understand the structure of the periodic table (see Fig. 73), the fact that when atoms are tabulated according to atomic number their properties repeat themselves in a regular way. Thus hydrogen has but one electron, which naturally prefers to go into the K shell. Of course, it does not have to; it can be in a higher shell, or level, corresponding to a higher energy, and then it is an excited electron. But this is not a stable situation: collision with another atom or molecule, or interaction with radiation, is most likely to absorb the extra energy and permit the atom to fall to its lowest and most stable energy level, losing its excitation, so that this lowest level is the normal state. This situation of the existence of excited states, but the preference for the normal state, is characteristic of all the atoms, and for the moment we are describing the normal states, for they are the ones in which we ordinarily find the atoms.

To resume, helium has two electrons, and in the normal state they are both in the K shell. This shell is now completed, no

more electrons can be bound in it, and such a completed shell is characteristic of the inert gases, of which helium is one. Lithium, with three electrons, would have two *K* electrons, and one *L*, and the latter would be loosely bound, and could be easily detached. In connection with this, we observe that lithium is an alkali metal, very much inclined to form a singly charged positive ion, which it does by losing the one electron, the loss of unit negative

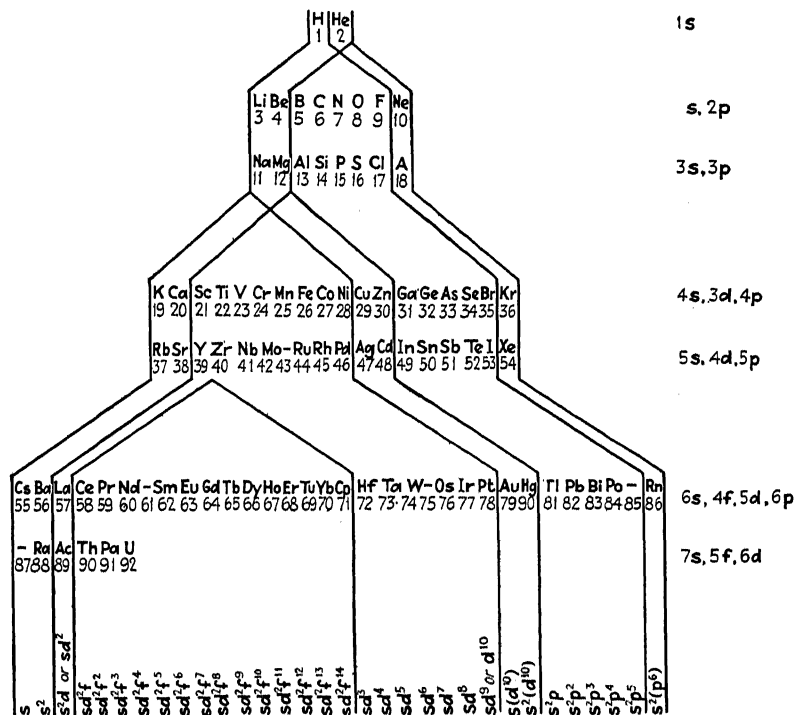


FIG. 73.—Periodic table of the elements, with electron configuration of lowest states.

charge being the same as gaining unit positive charge. Next, beryllium with four electrons has two *K*'s and two *L*'s, and can easily lose the latter to form a divalent positive ion. Thus we go through boron with two *K*'s and three *L*'s, carbon with four *L*'s (forming sometimes the ion with four positive charges) and nitrogen with five *L*'s. By this time, however, the attractions between the outer electrons and the nucleus have become rather large, and they are not easy to detach. The reason for this is that as we get more electrons in a shell, the effective nuclear

charge gets larger. For the electrons in a shell cannot shield each other very effectively; off hand we cannot say whether they are inside or outside the sphere of the last paragraph, and as a matter of fact the contribution to the shielding constant made by an electron in the same shell we are considering is only about 0.35 of an electronic unit. Thus if the effective nuclear charge for lithium's L electron were 1.30 (which is about the right amount, equal to $Z - S$ where $Z = 3$, $S = 1.70$ for the two K electrons, which do not shield perfectly), then for one of the two L electrons in beryllium we should have $4.00 - 1.70 - 0.35 = 1.95$, and for an L electron in boron $5.00 - 1.70 - 0.70 = 2.60$, increasing 0.65 for each atom, until for nitrogen we have 3.90 and for oxygen 4.55. Since the electrostatic attractions are proportional to the nuclear charge, this means that it is much harder to remove an electron from nitrogen than from lithium. By the time we come to oxygen and fluorine, we hardly have positive ions formed at all. But now another situation comes in: the attractions become so strong that an atom can pull an extra electron or two into its outer shell, forming a negative ion. Thus oxygen very easily forms a singly charged negative ion, and sometimes a doubly charged one. It can not go farther than this, for with two extra electrons its L shell has eight electrons and is completed. Similarly, fluorine can form a singly charged negative ion, but no more. And finally neon, with ten electrons, has two K 's, eight L 's, and consists of closed shells. It is the next inert gas after helium. It forms no ions: it would have to hold an extra electron in the M shell, and this would not be tightly bound, so that it would not stay; or to form a positive ion, it would have to lose one of its L electrons, and these are held too tightly to be removed by ordinary chemical processes. Thus it is inert.

After neon, we next come to sodium, with eleven electrons. This has two K 's, eight L 's, and the next electron must be an M . That is, it has one loosely bound electron, just like lithium. It again has a tendency to form a singly charged positive ion, and is an alkali metal like lithium. Magnesium, next, has two M electrons, and is like beryllium. We begin here to see the origin of the periodic table, for we have advanced by eight in our series of elements and have come to elements of similar properties. The similarity persists in this way up through argon, with eighteen electrons. At that point, we must take account of a further fact which we have not mentioned. Each of these shells is

really subdivided into subshells, of slightly different size and energy. The subshells are determined by the azimuthal quantum numbers, the states s , p , d , . . . of the same total quantum number becoming less tightly bound as we go out in the series, on account of decreased penetration. The maximum number of electrons in a shell of a given designation is invariable: an s group can have only 2 electrons, a p group 6, a d group 10, an f group 14, and so on (2×1 , 2×3 , 2×5 , 2×7 , . . . , or in general $2 \times$ the number of subgroups of different m values). Now the K shell contains only the s group, accounting, therefore, for its maximum number 2 of electrons. The L shell contains a $2s$ and a $2p$ group, so that its maximum number is $2 + 6 = 8$. Similarly the M has subshells $3s$, $3p$, $3d$, with a maximum number $2 + 6 + 10 = 18$, and N has $4s$, $4p$, $4d$, $4f$, with a possibility of $2 + 6 + 10 + 14 = 32$ electrons. When now we examine the energies of these various groups, we discover that the differences of energy between different subgroups of a shell may often be larger than those between different shells, with a result that the order of groups is changed. As a matter of fact, beginning with the most tightly bound shells, the groups are arranged as far as their energy is concerned approximately as shown in the following table, in which the first line gives the group, the second the number of electrons in the group, the third the total number of electrons in that group and all inside it, and the last the element completing the group, whose atomic number therefore stands just above it:

1s,	2s,	2p,	3s,	3p,	4s,	3d,	4p,	5s,	4d,	5p,	6s,	4f,	5d,	6p,	7s
2	2	6	2	6	2	10	6	2	10	6	2	14	10	6	2
2	4	10	12	18	20	30	36	38	48	54	56	70	80	86	88
He	Be	Ne	Mg	A	Ca	Zn	Kr	Sr	Cd	Xe	Ba	Yb	Hg	Rn	Ra

Within each shell the subshells are arranged in the order stated, but there is overlapping between the shells.

We now see that at A (argon), although the M shell is not completed, still the $3p$ subshell is, and this is enough to form a closed group and an inert gas. Next we come to K, 19, with one $4s$ electron, another alkali, and Ca, 20, with two, an alkaline earth like Be and Mg. But now instead of forming a group of 8 by adding p electrons, the next additions go into the $3d$ shell, and only after that is filled up do they go into $4p$, so that by the time we come to the next inert gas, Kr, we have added 18 electrons rather than 8 after A. The series of elements in which the $3d$

electrons are being added is the iron group. These have considerable similarity, because although the $3d$ electrons are less tightly bound than the $4s$, they are farther inside the atom, and the outside parts of these atoms are quite similar. When we go beyond Kr, we repeat the same sort of process, having another group of 18 elements in which the $5s$, $4d$, and $5p$ electrons are being added, before coming to the next inert gas Xe. The transition group which we go through here is the Pd group. Next after that, after adding the two $6s$ electrons to form Ba, the whole group of 14 $4f$ electrons is added, resulting in a long group of remarkably similar elements, the rare earths. As a matter of fact, these elements have one $5d$ electron each, so that our scheme is a little misleading in respect to them. After finishing the $4f$ group, the normal procedure repeats itself, the $5d$ and $6p$ being added to complete the shell of 18 interrupted by the rare earths and terminated at Rn, and finally the 7-quantum electrons being added to give the elements of the last, incomplete row of the table.

It is often convenient, in describing an atom in any state, to give the number of electrons having each quantum number by a symbol, as $1s^2 2s^2 2p^6 3s$ for the normal state of Na, meaning that there are two $1s$, two $2s$, six $2p$, one $3s$ electron. Such an arrangement is called a configuration. And a transition between two stationary states can be conveniently denoted by writing the two configurations. Thus the transition $1s^2 2s^2 2p^6 4p \rightarrow 1s^2 2s^2 2p^6 3s$ for Na is a line of the principal series in the optical spectrum; the transition $\text{Na } 1s^2 2s^2 2p^6 3s \rightarrow \text{Na}^+ 1s^2 2s^2 2p^6 3s$ represents the process of ionizing one of the K electrons of Na; and so on.

247. The Method of Self-consistent Fields.—We have just seen that the electrons of an atom act approximately as if they moved in central fields, rather than under the action of the other electrons, and have shown that this leads to quantum numbers for the electrons, to shells resulting from this, and to the periodic properties of the elements as successive shells are filled up. In making this idea more precise, we meet the method of self-consistent fields, developed by Hartree. In this method we assume that

1. The field in which the k th electron moves is obtained by taking the wave function of each of the other electrons, squaring to get the average density of charge due to these electrons, averaging over angles to get a spherically symmetrical distribution,

adding all these charge densities together, and finding the potential, together with that of the nucleus, by electrostatics. This, of course, will give a nonhydrogenic field, different for each electron.

2. To get the wave function of the k th electron, we solve Schrödinger's equation for the field above, using the appropriate quantum numbers. Since the field is nonhydrogenic, we must use numerical methods, or the Wentzel-Kramers-Brillouin method.

Having found these final wave functions, they must be the same ones with which we started step 1. It is this fact which leads to the name "self-consistent." If we started with arbitrary wave functions, computed a field, solved for the wave functions in that field, the final functions would not in general agree with the original ones. If we keep on repeating the process, however, using in each case the final wave functions of one stage of the calculation to begin the next, it rapidly converges so that after a few repetitions the field is approximately self-consistent. This method has been used for numerical computation of the wave functions of a number of atoms.

248. Effective Nuclear Charges.—The method of self-consistent fields, though quite accurate, demands numerical computation, and is not well suited for elementary calculations. We may instead approximate the wave function of each electron by a hydrogen wave function, corresponding to an effective nuclear charge $Z - S_i$. To get S_i , we should add up the total number of electrons within a sphere whose radius is the effective radius of the i th electron's wave function. It is easier to figure, not by means of the radius, but from the quantum number, since to a rough approximation the radius of an orbit is $n_i^2/(Z - S_i)$, so that electrons inside a given one are those of smaller total quantum number. The following table proves to give roughly the contribution to the shielding constant of a given electron from each other type of electron, valid for the electrons found in the light atoms. We see that the shielding of one electron by a second does not go suddenly from unity to zero as the shielding electron's quantum number becomes greater than that of the shielded electron, but instead changes gradually, in accordance with the fact that each electron really has charge distributed over all distances, and it is possible for part of the charge to be inside, part outside, a given radius.

TABLE 1.—CONTRIBUTION OF ONE SHIELDING ELECTRON, OF GIVEN QUANTUM NUMBER, TO SHIELDING CONSTANT OF SHIELDED ELECTRON

Shielded electron	Shielding electron				
	1s	2s	2p	3s	3p
1s	0.35	0	0	0	0
2s	0.85	0.35	0.35	0	0
2p	0.85	0.35	0.35	0	0
3s	1.00	0.85	0.85	0.35	0.35
3p	1.00	0.85	0.85	0.35	0.35

To illustrate the use of this table, let us take the case of Na, $Z = 11$, in its normal state $1s^2 2s^2 2p^6 3s$. Evidently we have three shells, corresponding to the three values of n . Then we have

$$n = 1: S = 0.35, \text{ radius} = n^2/(Z - S) = 1/10.65 = 0.09$$

$$n = 2: S = 2(0.85) + 7(0.35) = 4.15, \text{ radius} = 4/6.85 = 0.58$$

$$n = 3: S = 2 + 8(0.85) = 8.80, \text{ radius} = 9/2.20 = 4.09.$$

The inner radii are as given in Prob. 9, Chap. XXXIII.

The calculations we have given so far refer to wave functions, rather than energy levels. To investigate the latter, we must make a more careful discussion of the theory of the many-body problem and its treatment by Schrödinger's equation.

249. The Many-body Problem in Wave Mechanics.—Our treatment of atomic structure so far has been rather intuitive, not based directly on Schrödinger's equation at all. We have not yet set up the problem of many bodies in wave mechanics. To do so, we proceed as follows: Let the problem have N generalized coordinates, $q_1 \dots q_N$. Then we seek a wave function $\psi(q_1 \dots q_N, t)$, such that $\psi^2 dq_1 \dots dq_N$ gives the probability that the coordinates will be found at time t in the region $dq_1 \dots dq_N$. To set up Schrödinger's equation, we take the classical Hamiltonian function, convert it into an operator H by substituting $\frac{h}{2\pi i} \frac{\partial}{\partial q_i}$ for p_i , and write the equation $H\psi = -\frac{h}{2\pi i} \frac{\partial \psi}{\partial t}$. We eliminate time as usual, and have a differential equation for $u(q_1 \dots q_N)$, which is $Hu = Eu$, E being the energy of the whole system.

There is one simple case of the many-body problem: that where there are many particles, exerting no forces on each other.

That is, we may have n particles, whose coordinates are $x_1y_1z_1 \dots x_ny_nz_n$, and the potential is $V = V_1(x_1y_1z_1) + \dots + V_n(x_ny_nz_n)$, without any terms involving coordinates of two particles simultaneously. For such a potential, $\frac{\partial V}{\partial x_i} = \frac{\partial V_i}{\partial x_i}(x_iy_iz_i)$, a force on the i th particle depending only on the coordinates of that particle. In such a case, we can separate variables, writing $u = u_1(x_1y_1z_1) \dots u_n(x_ny_nz_n)$. For Schrödinger's equation can be written

$$\left[\left(-\frac{\hbar^2}{8\pi^2m_1} \nabla_1^2 + V_1 \right) + \dots + \left(-\frac{\hbar^2}{8\pi^2m_n} \nabla_n^2 + V_n \right) \right] u = Eu, \quad (1)$$

where ∇_i^2 means $\partial^2/\partial x_i^2 + \partial^2/\partial y_i^2 + \partial^2/\partial z_i^2$. A separation of variables can be carried through in the usual way, and can be summarized as follows: if we write u as a product, as above, then Schrödinger's equation is satisfied if

$$\left(-\frac{\hbar^2}{8\pi^2m_i} \nabla_i^2 + V_i \right) u_i = E_i u_i, \quad (2)$$

$$E_1 + \dots + E_n = E$$

In the case of atomic structure, and in general with the structure of matter, there are forces between the electrons. But here it is possible to make an approximation, as we have done: we replace the actual force between a given electron, say the i th, and the others, by the average which it would have from the mean distributions of the other electrons in space. Roughly we may say that, while the force with any particular arrangement of the other electrons will differ from this value, it will average out to give our mean value, and the deviations from the mean will not be so large as to destroy the approximation. Thus, using such a method, each electron becomes acted on, not by the other electrons, but by an averaged field. It is the motion in this field that we have considered in the present chapter.

250. Schrödinger's Equation and Effective Nuclear Charges.—The result of the approximate calculation we have made has been a set of one-electron wave functions, one for each electron of the atom. These satisfy equations which, in atomic units, are

$$\left[-\nabla_i^2 - \frac{2(Z - S_i)}{r_i} \right] u_i = -\frac{(Z - S_i)^2}{n_i^2} u_i. \quad (3)$$

Now the potential energy of the whole atom, in atomic units, is

$$-\sum_i \frac{2Z}{r_i} + \sum_{\text{all pairs}} \frac{2}{r_{ij}}, \quad (4)$$

if r_{ij} is the distance between the i th and j th electrons. Thus the Hamiltonian is

$$H = \sum_i \left(-\nabla_i^2 - \frac{2Z}{r_i} + \sum_{j \text{ inside } i} \frac{2}{r_{ij}} + \sum_{\substack{j \text{ in same} \\ \text{shell as } i}} \frac{1}{r_{ij}} \right), \quad (5)$$

where the two summations are the same thing as the sum over all pairs. If now we assume that $u = u_1 \cdots u_n$, where the u 's are as we have found, and try to see how good an approximation this forms, we have, substituting for the Laplacians, from Eq. (3),

$$Hu = \left[-\sum_i \frac{(Z - S_i)^2}{n_i^2} \right] u + \sum_i \left(-\frac{2S_i}{r_i} + \sum_{j \text{ inside } i} \frac{2}{r_{ij}} + \sum_{\substack{j \text{ in same} \\ \text{shell as } i}} \frac{1}{r_{ij}} \right) u. \quad (6)$$

If Schrödinger's equation were satisfied, this would be Eu , where E is a constant. This is not true; the first term is a constant times u , but the second is a variable function of the r 's times u . The average value of the last term, however, is approximately zero. For $2/r_{ij}$ is the potential, at the i th electron, of the j th electron. If the latter is inside, and we average over its position, and average to make it spherically symmetrical, the potential will be the same as if it were concentrated at the center, or will be $2/r_i$. For an electron in the same shell, it turns out that the average of $1/r_{ij}$ is about $2(0.35)/r_i$. The summation, for all electrons inside or in the same shell as i , is then essentially $2S_i/r_i$, just canceling the first term, and leaving as the result, using this approximate method of averaging, of

$$Hu = -\sum_i \frac{(Z - S_i)^2}{n_i^2} u,$$

showing that we have an approximate solution, and that the energy of the atom is $-\sum_i \frac{(Z - S_i)^2}{n_i^2}$. This represents the nega-

tive of the energy required to remove all the electrons from the atom. If we wish to find the energy of the atom by first-order perturbation theory, we recall that we must find the diagonal term of the energy matrix, or $\int \psi^* H \psi dv$. This means averaging the energy over the wave function, or over the motions of the electrons; and to the same approximation we have just used, the summations average to zero, leaving the same energy we just found.

As an example of the calculation of energy, we can again take the case of Na. The energy of normal Na is, using the shielding constants found above, $- \left[2 \left(\frac{10.65}{1} \right)^2 + 8 \left(\frac{6.85}{2} \right)^2 + \left(\frac{2.20}{3} \right)^2 \right] = -321.4$ units. With one 1s electron removed, making the appropriate changes in shielding constants, the energy is $- \left[\left(\frac{11.00}{1} \right)^2 + 8 \left(\frac{7.70}{2} \right)^2 + \left(\frac{3.20}{3} \right)^2 \right] = -240.6$ units. The difference is 80.8 units, or 1,094 volt-electrons, representing the ionization potential. Similarly with the 2s removed, the energy is $- \left[2 \left(\frac{10.65}{1} \right)^2 + 7 \left(\frac{7.20}{2} \right)^2 + \left(\frac{3.05}{3} \right)^2 \right] = -318.6$, leaving an ionization potential of 2.8 units, or about 38 volt-electrons. Finally the ionization potential of the 3s, as we immediately see, is simply $\left(\frac{2.20}{3} \right)^2 = 0.54$ unit = 7.3 volt-electrons.

251. Ionization Potentials and One-electron Energies.—In the method of self-consistent fields, each electronic wave function is the solution of a central field problem, for a single electron. This one-electron problem has a certain energy, as found in the preceding chapter, always negative, very large numerically if the electron is tightly bound, smaller if it is more loosely bound, and it is natural to ask for the interpretation of this energy. The connection with tightness of binding suggests directly that the one-electron energies measure the work required to remove the electron in question, or the ionization potential, the negative energies being the negative of the ionization potentials. This proves in fact to be the case. One can compute these ionization potentials, by finding the energies of the atom and ion and subtracting, and the result proves to be, to the first order of perturbation, just the one-electron energy. Thus the *K*

ionization potential is given by the distance of the 1s energy level below zero in the corresponding one-electron problem, and so on. The connection is not very accurate, but it is close enough to be very useful.

Our method of effective nuclear charges, being an approximation to the method of self-consistent fields, should show the same property, and we can give a simple though not entirely satisfactory proof. The negative of the ionization potential is the energy of the atom, minus the energy of the ion. If the i th electron is to be removed, and if S_j represents a shielding constant in the atom, S_j' for the ion, then the energy of the atom, minus the energy of the ion, is

$$-\sum_j \frac{(Z - S_j)^2}{n_j^2} + \sum_{j \neq i} \frac{(Z - S_j')^2}{n_j^2}. \quad (7)$$

If we set $S_j' = S_j - (S_j - S_j')$, and expand, this is

$$-\frac{(Z - S_i)^2}{n_i^2} + \sum_{j \neq i} \frac{(Z - S_j)^2 + 2(Z - S_j)(S_j - S_j') + (S_j - S_j')^2 - (Z - S_j)^2}{n_j^2}.$$

Our simple proof holds only in case there is no other electron in the same shell as the i th, and if we assume that each electron shields by either 1 or 0. Then we have $S_j - S_j' = 0$ if the j th electron is inside the i th, 1 if the j th is outside the i th. Thus for the ionization energy we have

$$-\frac{(Z - S_i)^2}{n_i^2} + \sum_{j \text{ outside } i} \frac{2(Z - S_i + \frac{1}{2})}{n_j^2}. \quad (8)$$

In this case we can easily find the energy of the one-electron problem. The potential energy of the field in which the larger part of the i th wave function is located is

$$-\frac{2(Z - S_i)}{r} + \sum_{j \text{ outside } i} \frac{2}{r_j}, \quad (9)$$

where S_i represents the number of electrons inside the i th, and the summation is for all outer electrons, assuming constant mean radii, which are approximated $1/r_j = (Z - S_j)/n_j^2$. To verify the correctness of this potential, we note that the corresponding force, $-[2(Z - S_i)]/r^2$, is what we should have for the charge

inside the sphere concentrated at the nucleus, and the constant terms of the summation are added to make the potential continuous at the outer shells. Thus the wave equation for u_i is

$$\left[-\nabla_i^2 - \frac{2(Z - S_i)}{r} + \sum_{j \text{ outside } i} \frac{2}{r_j} - E_i \right] u_i = 0,$$

which gives immediately

$$E_i - \sum_{j \text{ outside } i} \frac{2}{r_j} = -\frac{(Z - S_i)^2}{n_i^2},$$

or, using the value of $1/r_j$,

$$E_i = -\frac{(Z - S_i)^2}{n_i^2} + \sum_{j \text{ outside } i} \frac{2(Z - S_j)}{n_j^2}, \quad (10)$$

agreeing with the value (8) already found, except that the correction term $\frac{1}{2}$ in $(Z - S_j + \frac{1}{2})$ is missing. This formula, moreover, is interesting, in that it shows that the shielding has two effects on the energy: (1) The energy has the term $-(Z - S_i)^2/n_i^2$ instead of $-Z^2/n^2$, as we should have with an electron in the unshielded field of the nucleus. This effect, reducing the magnitude of the ionization potential, is called the inner shielding, since it comes from the inner electrons. (2) There is also the summation over the outer electrons, likewise resulting in a reduction of ionization potential, and called the outer shielding. As we see from our derivation, the outer shielding results from the rearrangement of shielding constants of the outer electrons when an inner electron is removed.

Problems

1. The K series in the x-ray spectra comes when a K electron is knocked out, and an L , M , . . . electron falls into the vacant place in the K shell. The lines are K_α (if an L electron falls in), K_β (an M), etc. Write down the configurations before and after the K_α and K_β transitions of Mo.

2. Show that the frequencies of the lines of the K series are less than the frequency of light necessary to cause ionization of the K electron. Compute the K ionization potential and the K_α line for Ca, and show that they fit in with the general case.

3. Moseley's law is that the square roots of x-ray term values (ionization potentials) form a linear function of the atomic number. This would obviously be true if there were just inner shielding, for then the square root would be simply $(Z - S)/n$. Investigate how closely this is true when there

is outer shielding as well, computing K and L term values for electrons from $Z = 10$ to $Z = 20$, and seeing how closely the square roots fall on straight lines.

4. Iso-electronic sequences are sets of ions, all of the same number of electrons, but with different nuclear charges, and hence different degrees of ionization. Compute the ionization potentials, or term values, $1s^2 2s^2 2p \rightarrow 1s^2 2s^2$, $1s^2 2s^2 3s \rightarrow 1s^2 2s^2$, for the atoms $Z = 5$ to 10 , indicating what ions they are (as $Z = 6$, $1s^2 2s^2 2p$ is C^+). Investigate to see whether these term values follow Moseley's law that the square root of the term value is a linear function of atomic number.

5. Using the approximation that the radius of a shell is $n^2/(Z - S)$, draw curves giving the radius of each shell as function of Z for all atoms up to $Z = 20$ (compute only enough values to draw the curves).

6. In a closed shell of p electrons, there are two electrons of $m = 1$, two of $m = 0$, two of $m = -1$. Using the spherical harmonics for these cases, compute the squares of the wave functions, treating these as electron densities. Add the densities of all electrons, showing that the sum is independent of angle, or that the p shell is spherically symmetrical. The same thing is also true of any completed shell.

7. Given a spherical distribution of charge, where the potential is $2Z_p/r$, and the force $2Z_f/r^2$, where Z_p, Z_f are both functions of r , prove that $Z_f = Z_p - r(dZ_p/dr)$.

8. Assuming that the electrons are located on the surfaces of spheres of radius $n^2/(Z - S)$, find and plot Z_f and Z_p for Na^+ as functions of r .

CHAPTER XXXV

INTERATOMIC FORCES AND MOLECULAR STRUCTURE

Atoms by themselves have only a few interesting properties: their spectra, their dielectric and magnetic properties, hardly any others. It is when they come into combination with each other that problems of real physical and chemical interest arise. Atoms act on each other with forces, in some cases attractive and in others repulsive, and in this chapter we shall consider the nature of these forces, how they arise, and what their results are in their effect on the physical and chemical structure of the substance. Interatomic forces in the first place hold atoms together to form molecules; this forms the province of chemistry. But in turn they hold molecules together into their various states of aggregation, as solids, liquids, and gases, and this is ordinarily considered to be part of physics. The distinction, however, is purely arbitrary, and not at all general. We shall begin by discussing the most important types of force, with a little consideration of the types of substances in which they are found. All the interatomic forces of interest in the structure of matter are electrical or in some cases magnetic; the only other forces, gravitational, are far too small to be of significance. We arrange the different types according to the way they depend on the distance of separation of the atoms.

252. Ionic Forces.—If two atoms are ionized, they attract or repel according to the inverse square. If the net charge on one is z_1 units, on the other z_2 , the potential energy between them is $z_1 z_2 e^2 / r$, if r is the distance between.

253. Polarization Force.—Atoms are polarizable, as we have seen in discussing refractive index, in Sec. 172, Chap. XXIV. That is, an atom in an electric field E acquires an electric moment αE . Now suppose that we have an atom or an ion in the presence of another ion. The ion produces a field ze/r^2 . This in turn polarizes the first atom or ion, producing a moment $\alpha ze/r^2$. The resulting dipole reacts back on the ion, attracting it with a force equal to the field of the dipole (equal to the moment of the

dipole times $2/r^3$) times the charge on the ion, or $-(2\alpha z^2 e^2/r^5)$. The potential of this force is $-(\alpha z^2 e^2/2r^4)$, giving always an attraction.

254. Van der Waals' Force.—Ionic and polarization forces are met only with ions. The forces observable at largest distances between neutral atoms or molecules, and hence of importance in the behavior of liquids and imperfect gases, are called Van der Waals' forces, on account of their appearance in Van der Waals' equation of state for imperfect gases. They arise as follows. An atom is generally spherically symmetrical and thus on the average has no external electric field. But this is only on the average; instantaneously it is not spherical, but the electrons are at arbitrary positions, and the result gives a dipole moment, averaging to zero, but instantaneously different from zero. This dipole polarizes a second atom or molecule. Thus the field of a dipole of moment μ is $\mu/r^3 \times$ function of angle. In the two special cases where the dipole points straight toward, or away from, the atom, the function of angle has the values ± 2 , respectively. In that case, the induced dipole in the second molecule is $\pm(2\alpha\mu/r^3)$. This produces a field back on the first, equal to $\pm(4\alpha\mu/r^6)$. The force by which it acts on the original dipole is equal to the rate of change of the field with r , times the dipole moment, times a function of angle which is ± 1 in the two cases considered, or $\left(\mp \frac{24\alpha\mu}{r^7}\right)(\pm\mu) = -\frac{24\alpha\mu^2}{r^7}$, with potential energy $-\frac{4\alpha\mu^2}{r^6}$. If we had considered all angles, we should have got a different constant, but in any case an attraction, $-\text{constant} \times \alpha\mu^2/r^6$.

To calculate the polarization and Van der Waals' forces, we should have to find α and μ . The calculations for these are difficult and will not be attempted here, though a derivation will be given in a later chapter. For the present, however, we can get some semiempirical formulas which will serve for rough calculations. First, the polarizability α has the dimensions of a volume. An argument from a simple model in Sec. 172 showed that, at least in order of magnitude, the polarizability of a spherical atom is equal to the cube of its radius. Now the radius of an electron's orbit can be approximated by $n^2 a_0 / (Z - S)$, so that we might imagine that the polarizability of an atom could be approximated by the sum of such terms, cubed, for all elec-

trons. Empirically, one finds that this gives about the right dependence on Z , but not very accurately for n : the contribution of an electron to the polarizability proves to be approximately

$$\left(\frac{n^2 a_0}{Z - S}\right)^3 \times \begin{cases} 4.5 & \text{if } n = 1 \\ 1.1 & \text{if } n = 2 \\ 0.65 & \text{if } n = 3, \text{ etc.} \end{cases}$$

The total polarizability is the sum of such contributions, for all electrons. As we readily see, only the electrons in the outer shell make an appreciable contribution, since they have the largest values of n and the smallest Z 's. Hence we may simply multiply the number ν of electrons in this shell by the term above. Thus $\nu = 2$ for an ion with the same structure as the He atom, 8 for one built like Ne, 8 for one like A, etc.

For the Van der Waals' force, we expect an energy — constant $\times \alpha\mu^2/r^6$. We shall consider the problem more in detail in Chap. XLII, Sec. 301, where it is shown that the energy is

$$-\frac{3}{2} \frac{1}{r^6} \alpha \mu^2,$$

and where in addition we have the relation

$$\mu^2 = \frac{\alpha \Delta E}{2}.$$

In this formula, ΔE is the difference of energy of that transition from the normal state which contributes most to the refractive index and dispersion. Ordinarily this can be taken to be the same as the ionization potential of the atom. Thus, since we know how to find ionization potentials from our effective nuclear charges, we may use empirical or approximately calculated polarizabilities to get coefficients for the Van der Waals' attraction.

The three types of force we have enumerated all fall off as inverse powers of the distance. If we inquire further, we find that there is a whole series of terms, in higher and higher inverse powers of r . Thus between ions we have a series commencing with terms in $1/r$ and $1/r^4$, between atoms commencing in $1/r^6$, but having higher terms arising from interaction of the induced dipoles of both atoms with each other, interaction of dipoles and charges with quadrupole moments, etc. The complete series would be difficult to evaluate. In addition to these forces, there are other quite different ones, coming when the atoms are so close that their charge distributions actually begin to overlap.

Since these distributions fall off exponentially with distance, as in hydrogen functions, these types of force all fall off exponentially and for that reason cannot be expanded in inverse powers of r at all (the exponential function possesses a singularity at infinity and so cannot be expanded in power series in $1/r$). The forces are sometimes grouped together, but we prefer to break them up into three classes.

255. Penetration or Coulomb Force.—As one atom penetrates another, there will be forces on account of pure electrostatics, even if the two atoms do not distort each other. Let the outer shell of each atom penetrate within that of the other (Fig. 74a). Then the part of each which penetrates the other finds itself in a field attracting it toward the nucleus of the other, since it is no

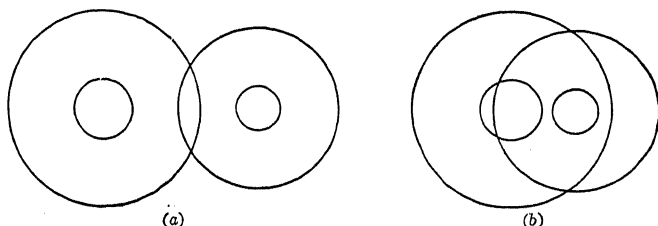


FIG. 74.—Penetration of one atom by another. Circles represent shells of electrons. (a) Attraction. Negative charge of each atom penetrates within the outer shell of the other, being attracted to the positive nucleus. (b) Repulsion. Nucleus of each atom penetrates the outer shell of the other, the repulsion of the nuclei for each other outbalancing the attractions.

longer shielded by all shells of the other. The result is an attraction of the charge of each for the other, pulling the whole atoms together. On the other hand, as the atoms get still closer, the whole system of inner shells of one would get inside the outer shell of the other (see Fig. 74b). These inner cores are both positively charged on the whole and will repel, a repulsion more than enough to counteract the attraction, in general. Hence at sufficiently close distance, the penetration force will be repulsive. In between, there will be some distance at which the force will be zero and there will be equilibrium.

256. Valence Attraction.—The penetration force acts even though the atoms are not distorted. The force of attraction principally concerned in valence, however, is an additional force resulting from the distortion of one atom by the other. The distortion produced by ordinary electrostatics is at least approximately taken care of by computing the polarization, as stated in

Sec. 253, but there is an additional effect, resulting from the operation of the exclusion principle, and the existence of electron spins, and which leads to a tendency for electrons to form stable pairs, agreeing with the ideas of G. N. Lewis regarding homopolar valence, or valence attraction between uncharged atoms. To understand this, even approximately, we must look more closely into the exclusion principle. In addition to their charge, electrons also act like little magnets, having a north and south pole. This is as if the charge were to rotate, forming a little electric current around a circle, and corresponding magnetic lines of force. The result is called electron spin. Now when we have a pair of electrons, it turns out that their spins can be oriented in just two possible ways: either parallel to each other, or opposite or antiparallel. If they are parallel, then the exclusion principle comes in and says they cannot be in the same shell. But if they are opposite the principle does not operate. It is a result of this that the allowed numbers of electrons in the various groups in an atom are all even numbers. Thus, in the *s* shell, after we have one electron, we can add a second if its spin is opposite to the first, for then the exclusion principle does not act. But if we now try to add a third, its spin must be parallel to one of the two already there, and the exclusion forbids it. Similarly a *p* group really contains three different subgroups, each of which can contain but two electrons, with opposite spins. Analogous results hold for the other groups. We see, then, that the subgroup of two electrons with opposite spins is a configuration which electrons like to form, and that only two electrons can enter such a configuration, so that there is a tendency toward pairing. But now it appears that such a pair can be formed by two electrons in different atoms, just as well as by two in the same atom. Thus if each of two atoms has just one electron, rather than two, in one of its subgroups, and if these two electrons have opposite spin, they can form a pair held in common by the two atoms, actually localized in the space between the atoms, and tending simply by electrostatic forces to hold the atoms together, the attractions of this negative concentration of charge for the nuclei, which must have a net amount of positive charge, more than counterbalancing the repulsions between like charges at large distances, though at smaller distances the force becomes repulsive, on account of the ordinary penetration effect. This is the origin of homopolar valence. We see that every electron

lacking from a closed shell can be interpreted as giving the possibility of forming a valence bond, so that for example the halogens have a single valence, oxygen and sulphur have two, hydrogen has one (one electron missing from $1s$), and so on.

257. Atomic Repulsions.—If one brings two atoms close enough together, they will always repel and resist further approach. This is what we know physically as the impenetrability of matter. It is a result of the exclusion principle, again. If we force two atoms so close together that the shells of the two atoms overlap, and if these shells are all filled with electrons, then we are really trying to force more electrons into the same region of space than the exclusion principle allows. What happens is that the electrons then move outside of this region, the atoms become distorted, and the resulting increase of energy is interpreted as a force of repulsion between the atoms. These actions commence as soon as closed shells begin to overlap appreciably, and as a result the atoms have rather sharp boundaries, and for some purposes may be considered as having definite sizes. We should notice that, if the outer shells of the atoms are not closed, this repulsion can be altered. Thus, if two lithium atoms approach, each having a closed K shell but only one electron in its $2s$ shell, either of two things can happen. If the two L electrons happen to have parallel spin, then the exclusion principle operates between them, and they will repel each other, as if they had only closed shells. But if the spins are opposite, then the outer shells can coalesce, forming a shared electron pair, and resulting in attraction. Even in such a case, however, we finally meet repulsion as we bring the atoms together. In the first place, at close enough separation, the K shells would begin to overlap, and since they are closed shells they would repel in the usual way. But also the pure electrostatic interaction gives repulsion at small enough distances. For with more and more penetration, we get to the point where the nuclei are close together, in the midst of a combined set of shells of electrons from both atoms. Increasing closeness will then increase the repulsion between the nuclei, without much changing anything else, and this repulsion will finally become great enough to cancel all other effects.

258. Analytical Formulas for Valence and Repulsive Forces.—

The three types of force which we have just been discussing, Coulomb penetration force, valence attraction, and repulsion,

all depend on the actual overlapping of the charge distributions of two atoms. Here again we can find a simple approximate formula, which is yet accurate enough to be decidedly useful. Since the charge distribution falls off in general exponentially with the distance, we may assume that the potential energy also falls off exponentially: energy = Ce^{-ar} , where r is the distance between nuclei. The constant C is negative for attractions, positive for repulsions. The value of a , of course, will be different with each type of force, and each type of atom. Nevertheless, we can give extremely rough rules which yet suffice to give the order of magnitude of a . First we set up, for each of our two atoms, the "radius" of the outer shell, $n^2/(Z - S)$. We add these radii for the two, multiply by 1 if the electrons in the outer shells are p electrons, as in closed shells, but by 1.4 if they are s electrons in both atoms, as in a molecule made of two alkali atoms. Let the result be r_0 . Then as far as order of magnitude is concerned, the energy is a constant times $e^{-4(r/r_0)}$ for the pure repulsion between closed shells. In the valence attraction case, where the curve has a minimum, we can combine the valence and Coulomb forces, since both behave about the same. Then the result is approximately

$$Ce^{-6(r/r_0)} - C'e^{-3(r/r_0)},$$

the first term representing the repulsion close in, the second the attraction farther out. The constants as we have written them are for the normal state of the atoms and molecules, and in this case it is found that the equilibrium distance for the valence attraction comes approximately at r_0 . This results, as we readily verify, by writing the formula in the form

$$D\left[e^{-6\left(\frac{r}{r_0}-1\right)} - 2e^{-3\left(\frac{r}{r_0}-1\right)}\right],$$

or more generally

$$De^{-2a(r-r_0)} - 2De^{-a(r-r_0)}, \quad (1)$$

where a is a constant, which we have set approximately equal to $3/r_0$. This form of potential curve has been used by Morse, and he has tabulated values of D , a , and r_0 for a number of molecules, in excited as well as normal states.

The constant coefficient D , or the corresponding coefficient in the pure repulsive energy, is not easily given in a general way. We can easily see its significance, however. In Fig. 75 we plot a Morse potential curve, observing that it has a minimum at r_0 ,

the energy at this point being $-D$, while at infinite separation the energy is zero. Thus D represents the energy required to pull the atoms apart to infinity if they are initially at rest at the equilibrium distance, or, in other words, the energy of dissociation of the pair of atoms. These energies, for actual molecules, vary between a fraction of a volt-electron and several volt-electrons, depending on the tightness of binding of the molecule.

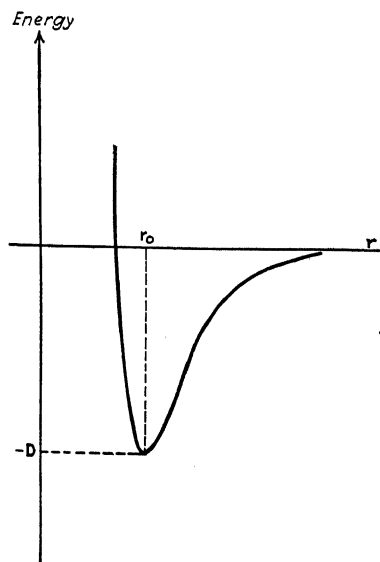


FIG. 75.—Morse potential curve,
 $De^{-2a(r-r_0)} - 2De^{-a(r-r_0)}$.

A few simple rules help in estimating D , as for instance that the larger r_0 , the smaller D tends to be (for example, F_2 is more tightly bound than I_2 , the F atom being smaller than I); molecules with a double or triple valence bond have larger D 's than with single bonds; etc.

The repulsive energy between closed shells, which we have approximated by $Ce^{-4(r/r_0)}$, is generally associated with an ionic or Van der Waals' attraction, resulting again in a minimum. This minimum, however, is ordinarily at much larger distances than r_0 , more nearly $2r_0$ or even larger. This is in consequence of two things:

the attractive forces are rather weaker than the valence attractions, and second the repulsion between closed shells is naturally larger, and effective at larger distances, than the repulsion found in valence compounds. In an actual case, where we know the Van der Waals' or ionic force, we can then make an estimate of the distance of separation at the minimum, and find C from the condition that the correct total potential has zero slope at this point. To get a number comparable with those met in valence attraction, we should write the repulsion in the form $De^{-4(\frac{r}{r_0}-1)}$. Then in actual cases D comes out of the order of a few volt-electrons.

Often one finds the repulsive forces of which we have just spoken approximated by an inverse power of r , as b/r^n , where n

proves to be about 8 or 9. We immediately see that both functions, exponential and inverse power, behave similarly, being large for small r , small for large r , so that either form can be used, though, since the repulsion depends on penetration, which actually goes off exponentially, we can be sure that the inverse power term is not so accurate. We can readily find out why n has about the value 9. The repulsive term is of importance, and can be found experimentally, and n determined, near the minimum of the energy curve. For Van der Waals' or ionic forces, as we have mentioned, this proves to come at about $2r_0$. Then suppose that we choose b and n so that b/r^n has the same value and slope as $Ce^{-4(r/r_0)}$ when $r = 2r_0$. We have

$$\frac{b}{(2r_0)^n} = Ce^{-4\left(\frac{2r_0}{r_0}\right)}$$

and

$$-\frac{nb}{(2r_0)^{n+1}} = -\frac{4}{r_0}Ce^{-4\left(\frac{2r_0}{r_0}\right)},$$

from which, dividing one by the other, $2r_0/n = r_0/4$, $n = 8$, approximately as is found experimentally. Many discussions, particularly of the structure of ionic crystals, are based on this inverse power formula, which has been used by Born and others.

259. Types of Substances: Valence Compounds.—Now that we have investigated the types of interatomic forces, we should consider them with reference to the different types of substances in which they occur. Broadly speaking, there are two main types of substances, corresponding to the two principal kinds of interatomic attractions, the ionic and the valence forces. Let us arrange our valence compounds roughly in order of melting or boiling points, starting with the most volatile, and ending with the most stable. The first substances on the list are not compounds at all, and indicate valence only in a sort of negative way: they are the inert gases, He, Ne, Ar, Kr, Xe. Since the outer shells of these are already completed, they form no ions, and they have no electrons to be shared and have no possibility of valence forces, and form no compounds. Next we come to a group of diatomic molecules, for example H_2 , O_2 , N_2 , F_2 , Cl_2 , Br_2 , CO , HCl , HBr , etc. These are held together by valence forces (HCl and HBr are somewhat ambiguous, and might be considered to be ionic compounds; this ambiguity is met in almost all H compounds). For example, each atom in H_2 has one electron;

they share these, making a pair. In O_2 , each atom has six L electrons; but they share two pairs (a double bond). As we go on, we come next to fairly simple polyatomic molecules. We have water, ammonia, methane: H_2O , NH_3 , CH_4 , all rather plainly valence compounds (though the ambiguity of which we spoke previously makes an ionic interpretation possible as well), with each hydrogen held by a single valence bond to the other atom. We might well include with these the ammonium ion, NH_4^+ , presumably built like methane. Other simple ones are CO_2 , CS_2 , with double bonds. Then we certainly should include some of the simple organic compounds, as acetylene C_2H_2 (triple bond between the carbons), ethylene C_2H_4 (double bond between the carbons), ethane C_2H_6 (single bond).

All these molecules of which we have spoken are held together by valence forces. On the other hand, there are also Van der Waals' forces between molecules, though of a smaller order of magnitude than the valence forces, and these hold the substances together in liquids and solids, all of low boiling points, but of increasing stability as the molecules become heavier and more complicated. The very considerable difference in order of magnitude between the valence and the Van der Waals' forces is significant, for this brings it about that the separate molecules preserve their identity, even when crowded close together.

More complicated organic compounds naturally come next in the list. They still preserve to some extent the property of existing as separate molecules, in gas, liquid, and solid, so that they still have both valence forces between atoms, and Van der Waals' forces between molecules. But as the molecules get more and more complicated, the Van der Waals' forces get larger and larger proportionally, so that with the fairly complicated ones they are of the same order of magnitude as the valence forces. Many complicated organic compounds dissociate when heated, rather than going through a change of state, since the heat necessary to melt and boil the substances becomes more and more nearly equal to that required to break up the molecules. It becomes, in other words, harder and harder to distinguish separate molecules, the solid acting more and more like a single big molecule.

The silicates form a group of compounds slightly suggesting the organic compounds in their complexity. They contain the group SiO_4^{-4} , which can be best described as a pure valence

compound, $\text{Si}(\text{O}^{-1})_4$, held together just like methane, Si being analogous to C. In many compounds the silicate groups are joined together, by sharing oxygens, as in the double group $\text{Si}_2\text{O}_7^{-6}$, or $(\text{O}^{-1})_3\text{Si}-\text{O}-\text{Si}(\text{O}^{-1})_3$, a neutral O atom being joined by its two bonds to the two Si atoms. This process of sharing oxygens may continue, until finally there is a network formed through the whole crystal, the metallic ions, as Ca^{++} , etc., merely fitting into empty space in the network, and all traces of molecular structure being lost. Thus these crystals are held together by forces so strong that they are not easily broken up. They are insoluble and refractory, and in fact form a great proportion of all the minerals.

260. Metals.—The metals form a type of substance more or less by themselves, but in general resembling valence compounds. There is a definite indication, at least in some of them, that there is a network of valence forces between the atoms, running through the metal, and holding it together to form a solid. At the same time, the simple Coulomb penetration force seems to account for a considerable part of the cohesion of metals. The network of valences seems to be connected with the electrical conductivity: an electron shared between two atoms can go to either one, and if the sharing exists through the solid, the electrons can migrate and carry a current. For many purposes, it is more correct in a metal to give up the idea that an electron is attached to a given atom at all, and treat them as free to move from one place to another, like the molecules of a perfect gas. The typical metallic states are solid and liquid. When a metal is vaporized, the tendency toward molecular formation does not seem to be strong. The vapors of such metals as have been examined show both monatomic and diatomic molecules; one wonders if polyatomic ones would not also be found if the experiment were made, acting simply like little pieces of the large metallic crystal.

261. Ionic Compounds.—The ionic compounds are not so easy to classify in a definite order as valence compounds, principally because they are more alike. The primary fact about ionic compounds is that they are held together by electrostatic forces, the atoms appearing in the ionized state. The forces between the atoms depend only on the distance, and are independent of the presence of other atoms (except in the matter of polarization). The laws governing the formation of ionic crystals are simple electrostatic ones, such as that positive and negative ions tend

to approach as closely as possible, ions of the same sign go as far apart as possible, charges in small volumes tend to equalize themselves, and so on. As a particular result of these, there is no tendency to form molecules. It is almost impossible to build up out of ions any structure which would not have large electrostatic fields around it; and further ions would be attracted by these fields, so that the substance can build up indefinitely. Further, the electrostatic fields are rather large, compared with the valence forces. The physical nature of these substances follows from the principles very easily. Their most characteristic form is the solid, where they form crystals in which the ions are arranged on a regular lattice. There is no trace of molecular structure in the lattice. They are hard and stable, often harder than metals, and of high melting point, although, of course, there is large variation from one compound to another. The vapor phase is an unimportant one for practically all ionic substances. Much more interesting in general than either liquid or vapor is the ionic state in water solution. Water, on account of its great dielectric constant, decreases all electrostatic forces. It thus almost removes the forces holding such a crystal together, and the solid breaks up into ions dissolved in the water.

When we ask about individual ionic compounds, we can well classify according to the ions from which they are made. The fundamental building stones are in every case ions of atoms; and the ions are of two sorts, positive and negative. The metals practically always form positive ions. They easily lose their valence electrons, as we have seen, so that all the electrons outside closed shells are removed, giving the alkali ions a charge 1, alkaline earths 2, the aluminum group 3, and so on. As we go through the series of elements, we see that even the nonmetals sometimes form positive ions, as Cl with seven positive charges. Sometimes, however, their ions are negative, though about the only important atoms forming negative ions are O and S, forming singly and doubly charged ions, and the halides F^- , Cl^- , Br^- , I^- . These atoms add electrons to make a closed shell, instead of losing them. It is obvious why there are so few: adding electrons makes an atom negatively charged, so that it tends to repel other electrons. It is a process which cannot go on far. The negative halide ions generally exist by themselves. The oxide ion also exists by itself in oxides; but it also forms complex negative ions, with positive nonmetallic ones, which are the most important negative ions

known. There are two alternative explanations of these radicals, either as pure ionic compounds, or as a combination of this with valence forces. For example, the sulphate ion can be regarded as being formed from a completely stripped sulphur ion and doubly charged oxygens: $\text{SO}_4^{-2} = \text{S}^{+6}(\text{O}^{-2})_4$. But if we assume that the oxygens have only single negative charges, we have the other possible structure $\text{S}^{+2}(\text{O}^{-1})_4$. With this structure, the sulphur has four electrons, as carbon does, and so has four homopolar valence bonds; and the oxygens have the same electron structure as halogens, with a single valence bond. Thus the sulphur can be bound to the four oxygens by valence bonds, assisting the electrostatic attraction, and the structure would have similarity to methane or carbon tetrachloride. This latter explanation seems to be nearer the truth, since it can be calculated that the work required to form the completely stripped positive ion in the ionic model would be much greater than the work necessary to form the other structure.

Problems

1. Find the potential energy between two helium atoms, using our approximate methods for calculating Van der Waals' and repulsive forces, and compare with the more accurate value

$$\left\{ 7.7e^{-2.43r/a_0} - \frac{0.68}{(r/a_0)^6} \right\} 10^{-10} \text{ ergs,}$$

where $a_0 = 0.53 \times 10^{-8}$ cm. The polarizability of helium is $1.43a_0^3$, and its ionization potential 1.80 Rh. Compare these with simple calculated values.

2. Using the potential of Prob. 1, compute the equilibrium distance of separation between two helium atoms, and find the energy of dissociation, in ergs, and volt-electrons. Compare the equilibrium distance with the mean distance in the liquid, which has a density of 0.14, assuming atoms to be spaced on a regular lattice, so that the mean distance will be $1/\sqrt[3]{n}$, if n is the number of atoms per cubic centimeter.

3. Find a radius of the helium atom for use in kinetic theory, assuming that two helium atoms at temperature 300° abs., with kinetic energy of $\frac{1}{2}kT$, collide head on. Find how close they come before they stop, and compare this molecular diameter with the distance r_0 .

4. Two energy levels of H_2 coincide with the lowest energy of the atoms at infinite separation, one an attractive level (corresponding to valence binding, with the spins of the two electrons opposed), and one repulsive (the spins being parallel, so that the exclusion principle operates). Plot the energies of both terms as functions of distance, deriving the exponents according to our approximate laws, and determining the scale from the fact that the energy of dissociation of the molecule is about 4.3 electron volts,

and that the energy of the repulsive term at the distance of molecular equilibrium is about 8 electron volts.

5. Compute by our approximate laws the distance of separation r_0 of the atoms in the normal states of the valence compounds given below, and compare with the experimental values tabulated:

Compound	r_0 (Ångströms)	Compound	r_0 (Ångströms)
C ₂	1.31	I ₂	2.66
CN	1.17	NO	1.15
CO	1.15	O ₂	1.21
H ₂	0.76	SiN	1.57

6. Compute by our approximation polarizabilities for the following ions, and compare with the experimental values tabulated:

Ion	$a \times 10^{24}$	Ion	$a \times 10^{24}$
O ⁻⁻	1.60	S ⁻⁻	5.91
F ⁻	0.868	Cl ⁻	3.33
Ne	0.398	A	1.67
Na ⁺	0.292	K ⁺	1.12
Mg ⁺⁺	0.173	Ca ⁺⁺	0.785

7. Compare the distance of separation of atoms in the metallic crystals tabulated below with the sum of the quantities $\frac{n^2}{Z - S}$ for the two atoms.

Metal	Distance, Ångströms
Na	3.72
K	4.50
Ca	4.97

8. Compute the interatomic potential energy for NaCl at large distance, assuming it is composed of Na⁺ and Cl⁻, so that there will be the ionic force, and at the same time a polarization force, the sodium polarizing the chlorine. Show that the polarization of sodium by chlorine can be neglected. Using the polarizabilities of Prob. 6, show that the potential energy is

$$\left[-\frac{27}{r/a_0} - \frac{302}{(r/a_0)^4} \right] \text{ electron volts.}$$

9. The observed interatomic distance in the NaCl molecule is 2.73 Ångströms. Compute the constants C and a in the repulsive potential Ce^{-ar} .

Find a by the rules we have used, and determine C so that the sum of the repulsive potential, and the attractive potential of Prob. 8, will have a minimum at the required distance.

10. Using the value of a found in the preceding problem, find the equivalent value of n in the repulsive potential b/r^n for the NaCl problem, seeing how nearly it equals 9.

CHAPTER XXXVI

EQUATION OF STATE OF GASES

In the preceding chapter, we have considered interatomic forces, and their effect in determining the nature of substances. When we begin to think more precisely of what we mean by the nature of substances, we conclude that the equation of state, and the closely related specific heat, are among the most important properties. We shall, therefore, take them up, giving necessarily enough thermodynamics and statistical mechanics to make calculations possible. Our investigations will be concerned with the thermal motion of the nuclei, moving under the interatomic potential which we have investigated. We shall naturally not be able to treat all sorts of substances; liquids, for instance, are so complicated that comparatively little progress has yet been made in understanding their properties. But gases and crystal-line solids both present features of simplification which we can make use of.

262. Gases, Liquids, and Solids.—Before passing to our analysis, let us consider what types of behavior we wish to explain. We can conveniently divide our discussion into gases, liquids, and solids. A monatomic gas, as an inert gas, is the simplest case: we have only to find its pressure, and total energy, as a function of volume and temperature, a task which can be carried out when we know the law of force between molecules. Gases of valence compounds, however, are more complicated. Their equation of state is not much harder to approximate than with monatomic gases, at least at low density, for on account of the rotation of the molecules they act on the average as if they were spherically symmetrical, and we need use only the intermolecular force averaged over angles in deriving the equation of state. In the specific heat, however, there are two forms of energy to consider: the translational kinetic energy of the molecules as a whole, which acts just as in monatomic gases, but also the rotational and vibrational energy of the individual molecules. This involves

a different sort of calculation. A still further complication appears in gases of some ionic substances, and of some valence compounds like I_2 and NO . Here there are several types of molecule which can be simultaneously present in the gas, as $2I \rightleftharpoons I_2$, $2Na \rightleftharpoons Na_2$, and a proper treatment of the equation of state and specific heat would demand investigation of the equilibrium concentrations of the constituents, and their change with pressure and temperature.

A liquid is more complicated than a gas, in that the molecules are so closely in contact that they can no longer be treated as points. The liquids of the inert gases are, of course, exceptions, and there are a few other exceptions, diatomic and polyatomic substances whose molecules rotate even in the liquid, and so act like spherical systems. But with most liquids the molecules are bulky enough so that they do not rotate, and are definitely non-spherical in their average behavior. In considering the equations of state, in particular the compressibility, one can no longer, as with a gas, neglect the change of volume of the molecules with change of pressure. As the molecules become larger and larger, as with complicated organic compounds, the distinction between forces within and forces between molecules becomes lost, and the whole liquid must be treated as a single complex, the volume being determined more and more definitely by the space required to pack the atoms together.

The state of close-packing of atoms which we have just mentioned is definitely reached with solids. In fact, with noncrystalline solids, there is no sharp distinction between the states, as glass for instance shows, solidifying perfectly continuously from the liquid. The solids with definite melting points are the crystals, which have a definite lattice arrangement of the atoms which is not met in the liquid. This regularity of arrangement is the simplifying feature which makes it possible to treat crystals theoretically. We can here commence our discussion with the state at absolute zero of temperature, where the atoms are at rest, and the whole crystal is in a position of equilibrium of the interatomic forces. The compressibility of such crystals can be fairly easily found from the forces, and this has been carried through particularly successfully for some of the ionic crystals. Then we can treat the crystal in thermal agitation by investigating the small oscillations of the atoms about their positions of equilibrium, using the method of normal coordinates. This

makes it possible to consider both equation of state and specific heat with fair ease and generality.

Out of all the group of topics which we have suggested, there are a few which can be treated theoretically fairly successfully. First, there is the equation of state of rare gases, or of polyatomic gases whose molecules rotate so as to be spherically symmetrical on the average. This is what we take up in the present chapter. Secondly, there is the specific heat of rotation and vibration of molecules. Thirdly, one can consider the equilibrium between different types of molecules in a gas, the question of chemical equilibrium. Fourthly, the equation of state and specific heat of crystalline solids can be investigated. As a preliminary to these, we must extend our treatment of statistical mechanics, which we have already considered slightly in Chap. XXX. We first follow out the ideas of classical statistics a little further, treat the equation of state of a gas by those methods, and then go to quantum statistics, asking what changes are introduced.

263. The Canonical Ensemble.—Following Chap. XXX, we consider a phase space; that is, a space in which each coordinate and each momentum of the system is plotted as a variable. Let the coordinates be $q_1 \dots q_n$, the momenta $p_1 \dots p_n$, the Hamiltonian function $H(q_1 \dots p_n)$. Then the phase space has $2n$ dimensions, and a point in this space represents a whole system (for instance a sample of gas). Next we set up an ensemble of points in this space, the number in the volume element $dq_1 \dots dp_n$ being proportional to $f(q_1 \dots p_n)dq_1 \dots dp_n$. We assume all points of the ensemble to be equally likely; that is, we assume that the probability that the coordinates and momenta of the system actually lie in the region $dq_1 \dots dp_n$ is proportional to the number of points of the ensemble in this region, or is proportional to $f dq_1 \dots dp_n$. Then to find the average of any function of the coordinates and momenta, as $F(q_1 \dots p_n)$, we multiply by f , integrate, and divide by the integral of f : $\bar{F} = \frac{\int F f dq_1 \dots dp_n}{\int f dq_1 \dots dp_n}$, as we saw in Chap. XXXI.

Now in particular we set up the canonical ensemble,

$$f(q_1 \dots p_n) = \text{constant } e^{-\frac{H(q_1 \dots p_n)}{kT}},$$

where T is the absolute temperature. This ensemble gives the probability that a system in thermal equilibrium at temperature

T will have its coordinates and momenta within given limits. The essential physical reason for this is the following: Suppose we have two systems, the first of coordinates and momenta $q_1 \dots q_n, p_1 \dots p_n$, the second $q_{n+1} \dots q_m, p_{n+1} \dots p_m$, with the separate Hamiltonian functions $H_1(q_1 \dots p_n)$, $H_2(q_{n+1} \dots p_m)$. Then physically we know that, if 1 and 2 are at the same temperature, and are then allowed to interact slightly, as by interchanging energy, it will be found that they are already in equilibrium with each other, and they already form a combined system in equilibrium at this temperature. This, in fact, is the definition of equality of temperature. But this is satisfied for the canonical ensemble. Thus if the separate systems are in equilibrium, their distribution functions are

$$f_1(q_1 \dots p_n) = \text{constant } e^{-\frac{H_1(q_1 \dots p_n)}{kT}}$$

$$f_2(q_{n+1} \dots p_m) = \text{constant } e^{-\frac{H_2(q_{n+1} \dots p_m)}{kT}}.$$

By the laws of probability, then, the probability that simultaneously the coordinates $q_1 \dots p_n$ will be in the range $dq_1 \dots dp_n$, and that $q_{n+1} \dots p_m$ will be in $dq_{n+1} \dots dp_m$, is proportional to

the product of these probabilities, or constant $e^{-\frac{(H_1+H_2)}{kT}} dq_1 \dots dp_n dq_{n+1} \dots dp_m$. But now suppose that the two systems are allowed to interact. The combined system will have an energy $H_1 + H_2 + H'$, where H' is a small interaction potential, depending perhaps on all coordinates and momenta, negligibly small compared with the separate energies (as for instance an interatomic force between the negligibly small number of molecules on the boundary between the two systems, which permits the flow of heat between them). Then according to the canonical ensemble, the distribution of the whole combined

system in thermal equilibrium should be constant $e^{-\frac{(H_1+H_2+H')}{kT}}$. But we observe that, except for the negligible energy H' , this is just the distribution before the interaction, so that the two systems were already in equilibrium before the interaction, and by definition are at the same temperature. This result is true only with the canonical ensemble, since it depends on the exponential form, adding exponents being equivalent to multiplying the functions.

Suppose we choose the constant in the definition of the canonical ensemble so that $\int f dq_1 \dots dp_n = 1$, and avoid having to

bother with the denominator in taking averages; this corresponds to normalizing a wave function. Further, let us write the con-

stant in the form $\frac{F}{h^n}$, so that we have

$$f(q_1 \cdots p_n) = \frac{e^{\frac{F-H}{kT}}}{h^n}$$

and

$$\int f(q_1 \cdots p_n) dq_1 \cdots dp_n = 1 = \frac{1}{h^n} \int e^{\frac{F-H}{kT}} dq_1 \cdots dp_n. \quad (1)$$

Here F is a quantity of the dimensions of energy, a function of T , chosen to make the constant have the correct value. Since $e^{F/kT}$ is dimensionless, and since the function f must have the dimensions of $1/(dq_1 \cdots dp_n)$, in order to make its integral dimensionless, we must multiply by a constant of these dimensions. We have chosen $1/h^n$, which has the correct dimensions, since h is of the dimensions of pq . It is a purely arbitrary matter that we have chosen this particular constant, since in all ordinary physical applications the constant drops out anyway, and it does not imply the introduction of quantum theory into classical questions. We shall later see, however, that it simplifies the comparison with quantum theory to have it there.

264. The Free Energy.—Let us take the factor $e^{F/kT}$ out of the integral above (it does not depend on the q 's and p 's), and divide through by it. Then we have

$$e^{-\frac{F}{kT}} = \frac{1}{h^n} \int e^{-\frac{H}{kT}} dq_1 \cdots dp_n. \quad (2)$$

The integral on the right is often called the integral of state (we shall later see cases where it degenerates to a sum, called the sum of state). It is fundamental in thermodynamic applications. The quantity F is the free energy, and we proceed to investigate its properties. We have seen that it depends on the temperature; but we must also observe that it depends on the volume. To see how this comes about, let us think about the Hamiltonian function H , in particular for a gas. We are considering only the nuclear motion, so that H includes the kinetic energy of the nuclei, and the potential energy of the interatomic forces, as

discussed in the last chapter. But it also includes another term, if the gas is in an enclosure: the repulsion of the wall. The molecules of the gas, as they strike the wall, are repelled, so violently that they never penetrate the wall. We may say approximately that the potential energy becomes rapidly infinite as any molecule approaches the wall, and is infinite if any molecule is outside it, so that $e^{-(H/kT)}$ is zero in that case, and there is no probability of finding one of the molecules outside. Now this term in the potential depends on the volume of the vessel, the rapid rise of potential coming at the edge of the volume, which is adjustable. Thus we have $H(q_1 \dots p_n, v)$, where v is the volume, so that the free energy, which depends on an integral of this quantity, also depends on v as well as T .

Let us investigate the rates of change of the free energy with respect to volume and temperature. We have

$$\frac{\partial}{\partial v} \left(e^{-\frac{F}{kT}} \right) = -\frac{1}{kT} \frac{\partial F}{\partial v} e^{-\frac{F}{kT}} = \frac{1}{h^n} \int -\frac{1}{kT} \frac{\partial H}{\partial v} e^{-\frac{H}{kT}} dq_1 \dots dp_n,$$

or

$$\frac{\partial F}{\partial v} = \overline{\frac{\partial H}{\partial v}},$$

where we remember the formula for finding the average of any quantity. Now consider a cylinder filled with gas, closed with a piston of unit area. If we decrease the volume, the increment of volume being $-dv$, which therefore equals numerically the displacement of the piston, and if the pressure, and therefore the force on the piston is p , we shall do the work $-p dv$ on the system. This will represent the increase in energy of the system, or dH . Hence we have $dH/dv = -p$. We may consider this relation as stating that p is the generalized force connected with a generalized coordinate v , and therefore equal to the negative derivative of the energy with respect to this coordinate. Performing the average, we then have

$$\left(\frac{\partial F}{\partial v} \right)_T = -p. \quad (3)$$

Next we can differentiate the free energy with respect to temperature. We have

$$\frac{\partial}{\partial T} \left(e^{-\frac{F}{kT}} \right) = \left(\frac{F}{kT^2} - \frac{1}{kT} \frac{\partial F}{\partial T} \right) e^{-\frac{F}{kT}} = \frac{1}{h^n} \int \frac{H}{kT^2} e^{-\frac{H}{kT}} dq_1 \cdots dp_n,$$

or

$$F - T \frac{\partial F}{\partial T} = \bar{H}.$$

If we define \bar{H} , the mean energy of all systems of the ensemble, as the internal energy E , we have

$$F - T \left(\frac{\partial F}{\partial T} \right)_v = E, \quad (4)$$

the familiar Gibbs-Helmholtz equation.

From Eqs. (3) and (4), we have

$$dF = -pdv - \frac{E - F}{T} dT. \quad (5)$$

Now let us define the entropy S by the equation

$$F = E - TS, \quad S = \frac{E - F}{T}. \quad (6)$$

Differentiating, this leads to $dE = dF + TdS + SdT$. Similarly, Eq. (5) becomes $dF = -pdv - SdT$. Combining, we are led to

$$dE = TdS - pdv. \quad (7)$$

Equation (7) is the fundamental equation of thermodynamics, which we have derived from statistical methods. For the first law of thermodynamics is $dE = dQ - pdv$, where dQ is the heat absorbed in a process, pdv is the work done by the system. And the second law of thermodynamics is that for a reversible change (as our change is, since we assume that the distribution is always given by a canonical ensemble, which means that it is always in equilibrium), the quantity dQ/T is a perfect differential, dS . Combining these statements, we have Eq. (7).

The specific heat can be found immediately by differentiating the energy with respect to temperature at constant volume. Using Eqs. (4) and (7), it is

$$C_v = \left(\frac{\partial E}{\partial T} \right)_v = T \left(\frac{\partial S}{\partial T} \right)_v = -T \left(\frac{\partial^2 F}{\partial T^2} \right)_v. \quad (8)$$

Thus we can find the specific heat, as well as the equation of state, by differentiating the free energy. This makes it a very useful function, and its calculation, by means of the integral of state, is the usual method of deriving information about physical properties of substances. Of course, we could derive the same information from the energy itself as a function of volume and temperature, but it is not quite so convenient to calculate.

265. Properties of Perfect Gases on Classical Theory.—Let us apply the method of the free energy to the calculation of the equation of state and specific heat of a perfect gas, on classical mechanics. Let there be N molecules, each of mass m , so that

$$H = \sum_{i=1}^N \frac{p_{xi}^2 + p_{yi}^2 + p_{zi}^2}{2m} + V,$$

where V , the potential energy, is zero so long as all molecules are within the volume v , but becomes infinite if even one molecule strays outside. Then we have

$$\int e^{-\frac{H}{kT}} dq_1 \cdots dp_N = \int_{-\infty}^{\infty} e^{-\frac{p_{x1}^2}{2mkT}} dp_{x1} \cdots \int_{-\infty}^{\infty} e^{-\frac{p_{zN}^2}{2mkT}} dp_{zN} \\ \int \cdots \int e^{-\frac{V}{kT}} dx_1 \cdots dz_N. \quad (9)$$

Now by direct integration each of the integrals over the p 's is simply $\sqrt{2\pi mkT}$. The integral over coordinates is the integral of unity over all regions where the coordinates are inside v , 0 over the outside, so that it is $\int \int \int_v dx_1 dy_1 dz_1 \cdots \int \int \int_v dx_N dy_N dz_N =$

v^N . Thus we have finally

$$e^{-\frac{F}{kT}} = \left(\frac{\sqrt{2\pi mkT}}{h} \right)^{3N} v^N, \quad (10)$$

a function of temperature and volume as it should be, and the free energy itself is

$$F = -3NkT \ln \frac{\sqrt{2\pi mkT}}{h} - NkT \ln v.$$

From this we have at once $p = NkT/v$, giving the ordinary law of perfect gases, and $C_v = \frac{3}{2}Nk$, likewise a well-known result.

266. Properties of Imperfect Gases on Classical Theory.—

Next let us consider an imperfect monatomic gas, such as an inert gas. This differs only in that there is an additional term in the Hamiltonian, a sum of interaction energies of each pair

of atoms: $H = \text{kinetic energy} + \sum_{\text{pairs } i,j} V_{ij} + \text{repulsion of walls,}$

where V_{ij} may be the sum of a Van der Waals attraction between the i th and j th atoms at large relative distances, and an exponential repulsion at small distance. We then have

$$e^{-\frac{F}{kT}} = \left(\frac{\sqrt{2\pi mkT}}{h} \right)^{3N} \int \int \int_v dx_1 dy_1 dz_1 \cdots \int \int \int_v dx_N dy_N dz_N e^{-\sum \frac{V_{ij}}{kT}}. \quad (11)$$

The integration over the coordinates can be carried out in steps. First we integrate over the coordinates of the N th molecule. The quantity $e^{-\sum \frac{V_{ij}}{kT}}$, can be factored: it is equal to

$$e^{-\frac{\Sigma'}{kT}} e^{-\sum_{i \neq N} \frac{V_{iN}}{kT}},$$

where Σ' represents all those pairs which do not include the N th molecule. The first factor then does not depend on the coordinates of the N th molecule, and may be taken outside the integration over its coordinates, leaving

$$\int \int \int_v e^{-\sum_i \frac{V_{iN}}{kT}} dx_N dy_N dz_N.$$

We rewrite this as

$$\int \int \int_v dx_N dy_N dz_N - \int \int \int_v \left(1 - e^{-\sum_i \frac{V_{iN}}{kT}} \right) dx_N dy_N dz_N = v - W, \quad (12)$$

the first term being simply the volume, the second term being an integral to be evaluated. To investigate W , let us imagine all the molecules except the N th being in definite positions of space. If the gas is rare, the chances are that they will be well separated from each other. Now if the point $x_N y_N z_N$ is far from any of these molecules, the interatomic potentials V_{iN} will all be small, and

the integrand will be practically $1 - e^0 = 0$. Thus we have contributions to this integral only from the immediate neighborhood of each molecule. If all are alike, each of these contributions will be equal to

$$w = \iiint \left(1 - e^{-\frac{V_{iN}}{kT}}\right) dx_N dy_N dz_N,$$

a quantity which, though it formally involves the index i , actually is independent of i . In fact, if we imagine the i th molecule to be located at the origin, and remember that V_{iN} is a function of r , the distance from the origin, we see at once that

$$w = \int_0^\infty 4\pi r^2 \left(1 - e^{-\frac{V(r)}{kT}}\right) dr, \quad (13)$$

where we integrate to infinity instead of to the boundary of the vessel because the integrand is so small for r 's larger than molecular dimensions that it makes no difference. In terms of this, we then have

$$W = (N - 1)w. \quad (14)$$

Now when we integrate over the coordinates of the $(N - 1)$ st particle, we have just the same situation over again, except that there are only $(N - 2)$ remaining molecules, and so on. Thus finally we have for the integral over coordinates

$$[v - (N - 1)w][v - (N - 2)w] \cdots v.$$

We can easily evaluate this product, by taking its logarithm, which is what we want anyway. This is

$$\sum_{s=0}^{N-1} \ln(v - sw) = \sum_{s=0}^{N-1} \ln v + \sum_{s=0}^{N-1} \ln(1 - sw/v).$$

The first term is $N \ln v$, which we should have for the perfect gas. For the second, we note that on account of the rarity of the gas, sw/v is always small compared with 1. Hence we have $\ln(1 - sw/v) = -sw/v$ approximately, and the sum is approximately equal to the integral with respect to s , or $\int_0^{N-1} -\frac{sw}{v} ds = -\frac{(N-1)^2 w}{2v}$. To this order, then, neglecting unity in comparison with N , we have

$$F = -3NkT \ln \frac{\sqrt{2\pi mkT}}{h} - NkT \ln v + \frac{N^2 kT w}{2v}. \quad (15)$$

We then have $p = \frac{NkT}{v} + \frac{N^2wkT}{2v^2} + \dots$. This is often written in the form

$$\frac{pv}{RT} = 1 + \frac{Nw}{2v} + \dots, \quad (16)$$

where $R = Nk$. This expression pv/RT is called the virial, and the coefficients of its expansion in inverse powers of the volume are called the virial coefficients, so that $Nw/2$ is the second virial coefficient. The results of experiments on imperfect gases are ordinarily given as tables of the virial coefficients as functions of temperature, and by the equation above we can compute the second coefficient, finding w if necessary by numerical integration from $V(r)$. In addition to the pressure, we can, of course, find the specific heat, and it immediately comes out the same as for the perfect gas. We must remember, however, the rotational and vibrational specific heats of the polyatomic gases, which must be added to the translational terms to get the total specific heat.

267. Van der Waals' Equation.—There is a limiting case in which we can compute w approximately. This is the case where the attractive part of $V(r)$ varies slowly with r , while the repulsive part varies so rapidly that it can be considered zero if r is greater than r_0 , infinite if r is less. This is what we should have if the molecules were rigid spheres of diameter r_0 , attracted by the Van der Waals' attraction. If we let $V_0(r)$ represent the attraction, we have

$$\begin{aligned} e^{-\frac{V(r)}{kT}} &= e^{-\frac{V_0(r)}{kT}} \text{ if } r > r_0, \\ &= 0 \text{ if } r < r_0. \end{aligned}$$

The integral then is

$$w = \int_0^{r_0} 4\pi r^2 (1 - 0) dr + \int_{r_0}^{\infty} 4\pi r^2 \left[1 - e^{-\frac{V_0(r)}{kT}} \right] dr.$$

The first term is simply $\frac{4}{3}\pi r_0^3$, the volume of a sphere of radius r_0 , or eight times the volume of the sphere of diameter r_0 which represents a molecule. In the second integral, we may expand the exponential as a power series, since V_0 is relatively small: it is $1 -$

$[1 - (V_0/kT) \cdot \cdot] = V_0/kT$. Thus this term is $\frac{1}{kT} \int_{r_0}^{\infty} 4\pi r^2 V_0 dr + \dots$. If, for instance, we have the type of Van der Waals'

force considered in the last chapter, we have $V_0 = -\beta/r^6$, where $\beta = \alpha\mu^2$. Then the term is $-(4\pi\beta/3r_0^3kT)$. In this case, the second virial coefficient becomes

$$\frac{Nw}{2} = \frac{N}{2} \left(\frac{4}{3}\pi r_0^3 \right) - \frac{2N\pi\beta}{3r_0^3kT} \dots,$$

the further terms being in higher inverse powers of T . We may write this

$$b - \frac{A}{RT},$$

where b is four times the volume of all molecules, $A = 2N^2\pi\beta/3r_0^3$. Actual gases have second virial coefficients which agree well with this formula. The pressure, in other words, is given by the result

$$\frac{pv}{RT} = 1 + \frac{b}{v} - \frac{A}{RTv} + \dots, \quad (17)$$

being greater than for a perfect gas for large T (the b/v term preponderating), and less for small T . Physically, at high temperature, the finite size of the molecules, given by b , decreases the apparent volume, which produces an increase of pressure; while at lower temperatures the attractions between molecules, given by A , pull the gas together and decrease the pressure.

There is a very well-known equation, Van der Waals' equation, for the pressure of an imperfect gas. This is

$$\left(p - \frac{A}{v^2} \right) (v - b) = RT \quad (18)$$

This differs from the equation of state of a perfect gas in two respects: in having the volume $(v - b)$ in place of v , as if the molecules took up space, and in having the pressure decreased by the amount A/v^2 . The arguments used to deduce the equation are not reliable, and it cannot be regarded as more than a very useful empirical formula. But as far as the second virial coefficient is concerned, it is correct. If we compute pv/RT from it, and expand in inverse powers of volume and temperature, we can at once show that the expansion is what we have already found, as far as the term in $1/v$, the values of b and A agreeing with those we have already given. The higher terms in the expansion,

however, do not agree with what we should get by correct calculation.

268. Quantum Statistics.—Distribution functions, and hence canonical ensembles, have a rather different meaning in quantum theory from what they have in classical mechanics. For on account of the uncertainty principle we can no longer specify both coordinates and momenta, and hence cannot give functions of the q 's and p 's. Instead, as we have seen, we deal with a wave function ψ , such that $\bar{\psi}\psi$ gives the probability of finding the system at a given point of space. We could set up the corresponding quantity in classical statistics: if $f(q_1 \dots p_n)$ is the ordinary distribution function, normalized so that its integral is unity, then $\int \dots \int f(q_1 \dots p_n) dp_1 \dots dp_n$ would give a function of the q 's, giving the probability of finding the system with given q 's. Thus we should have the correspondence

$$\int \dots \int f(q_1 \dots p_n) dp_1 \dots dp_n \sim \bar{\psi}\psi,$$

the two quantities agreeing at any rate in the limit of large quantum numbers, where classical and quantum theory approach each other.

It is not difficult to show that this correspondence holds, at least with one degree of freedom. First, we consider microcanonical ensembles, ensembles in which all systems have the same energy, but are distributed in phase as if they had started off at all arbitrary instants of time. In such a case, with one degree of freedom, the probability of finding a system in a given range of coordinates is proportional to the length of time a system would stay in that range, or is inversely as its velocity. But now the corresponding quantum ensemble is one in which all systems are in the same stationary state. And using the Wentzel-Kramers-Brillouin method, we have already seen, in Chap. XXIX, that $\bar{\psi}\psi$ is approximately proportional to $1/\sqrt{E - V}$, or inversely as the velocity, so that in this case we actually have the correspondence we desire between classical and quantum theories. The same thing can be shown with more than one degree of freedom.

Now any kind of classical ensemble which is independent of time can be made up of microcanonical ensembles; we may regard it as consisting of a certain distribution on each energy surface. The corresponding situation is a quantum state in which all stationary states are excited at once, represented by a wave

function $\sum_k c_k u_k e^{-\frac{2\pi i}{h} E_k t}$. The corresponding density, averaged

over the rapid time fluctuations, is $\sum_k \bar{c}_k \bar{c}_k u_k$, corresponding to

a fraction $\bar{c}_k \bar{c}_k$ of all the systems being in the k th stationary state, or belonging to the particular microcanonical ensemble having energy E_k . Let us see what is the classical ensemble corresponding to this combination. We may approximate it in the following way. Let us imagine the energy surfaces corresponding to the stationary states drawn in the classical phase space. Then let a fraction $\bar{c}_k \bar{c}_k$ of the systems of the classical ensemble be uniformly distributed through the region between the k th and $(k+1)$ st energy surfaces, rather than just on the energy surfaces. We do this to get a continuous function. Then evidently the density of points between the k th and $(k+1)$ st surfaces will be $\bar{c}_k \bar{c}_k$ divided by the volume of phase space between these surfaces. This volume, as we have seen, is h^n . Then we have the approximation

$$f(q_1 \dots p_n) \sim \frac{\bar{c}_k \bar{c}_k}{h^n} \text{ between } E_k \text{ and } E_{k+1}.$$

This gives a step-like function for f , which would approach continuity as the stationary states got closer together. Now it is plain how we are to set up a canonical ensemble: we are to set $\bar{c}_k \bar{c}_k$ proportional to $e^{-E_k/kT}$, and this will then give the right variation for f . Of course, our correspondence is not exact, but we assume that the quantum canonical ensemble is the exactly correct thing, the classical one the approximation to it. This is justified by the fact that we can give just the same argument for the canonical ensemble's representing thermal equilibrium in the quantum theory that we could in classical theory, and we know quantum theory to be the correct form in cases where it differs from classical theory.

Having the canonical ensemble in quantum theory, we can now proceed to the calculation of the free energy and equation of state as we did in classical theory. To get exact correspondence, we should set

$$f(q_1 \dots p_n) = \frac{e^{\frac{F-H}{kT}}}{h^n} = \frac{e^{\frac{F-E_k}{kT}}}{h^n} = \frac{\bar{c}_k \bar{c}_k}{h^n}.$$

Now the integral $\int \dots \int f dq_1 \dots dp_n$ goes over into a sum over all stationary states, multiplied by the volume of phase space associated with each stationary state, or h^n . Thus we have

$$\int \dots \int \frac{e^{\frac{F-H}{kT}}}{h^n} dq_1 \dots dp_n \sim h^n \sum_k \frac{\bar{c}_k c_k}{h^n} = 1 = \sum_k e^{\frac{-E_k}{kT}},$$

and finally

$$e^{\frac{F}{kT}} = \sum_k e^{\frac{-E_k}{kT}}. \quad (19)$$

In the case of degeneracy, where there are several stationary states of the same energy, the sum in Eq. (19) includes a term for each state, so that for an energy level with g states, we have g times the contribution from a single level of the same energy.

269. Quantum Theory of the Perfect Gas.—We have already shown the correspondence between the classical and quantum expressions for free energy, to the approximation to which the Wentzel-Kramers-Brillouin method is accurate. This shows us that, for both the perfect and imperfect gases, we may expect to find about the same equation of state and specific heat on both theories. The errors in the method are large only when the wave length is changing very rapidly, and this actually comes, in this problem, only when two molecules are in collision with each other, or are colliding with the walls. Accurate discussion shows that there are appreciable corrections to the classical equation of state introduced in this way for the lightest gases (which therefore have longest wave length for a given velocity), but even these are small, and difficult to discuss. It is easy, however, to carry through the exact solution of the quantum theory of the perfect gas, and this will suffice to show the general situation.

Let the gas be confined in a rectangular volume of sides A, B, C . Then the wave functions for single molecules satisfying the boundary conditions of being zero on the boundaries are $\sin \frac{p\pi x}{A}$

$\sin \frac{q\pi y}{B} \sin \frac{r\pi z}{C}$, where p, q, r are integers. A wave function for the whole gas can be built up from this by multiplying together functions for all the molecules, obtaining

$$u = \sin \frac{p_1 \pi x_1}{A} \cdot \cdot \cdot \sin \frac{r_N \pi x_N}{C}.$$

Substituting in Schrödinger's equation,

$$\left[-\frac{h^2}{8\pi^2 m} \left(\frac{\partial^2}{\partial x_1^2} + \cdot \cdot \cdot + \frac{\partial^2}{\partial x_N^2} \right) + V \right] u = Eu,$$

where V is the same potential of repulsion of the walls which we have considered before, we at once have

$$E = \frac{h^2}{8\pi^2 m} \left(\frac{p_1^2 \pi^2}{A^2} + \cdot \cdot \cdot + \frac{r_N^2 \pi^2}{C^2} \right). \quad (20)$$

To get all states, we must take all combinations of the integers $p_1 \cdot \cdot \cdot r_N$, each going from one to infinity. Thus we have

$$\begin{aligned} e^{-\frac{E}{kT}} &= \sum_{p_1=1}^{\infty} \cdot \cdot \cdot \sum_{r_N=1}^{\infty} e^{-\frac{h^2}{8mkT} \left(\frac{p_1^2}{A^2} + \cdot \cdot \cdot + \frac{r_N^2}{C^2} \right)} \\ &= \sum_{p_1=1}^{\infty} e^{-\frac{h^2 p_1^2}{8A^2 mkT}} \cdot \cdot \cdot \sum_{r_N=1}^{\infty} e^{-\frac{h^2 r_N^2}{8C^2 mkT}}. \end{aligned} \quad (21)$$

Now at reasonably high temperatures, T is so large that we have to go to large values of the integers p , etc., before the exponential begins to fall off appreciably. Thus the terms of our summation differ only slightly from each other, and we can replace them by an integral, one factor being

$$\int_0^{\infty} e^{-\frac{h^2 p^2}{8A^2 mkT}} dp = A \frac{\sqrt{2\pi mkT}}{h}.$$

Thus we have

$$e^{-\frac{E}{kT}} = (ABC)^N \left(\frac{\sqrt{2\pi mkT}}{h} \right)^{3N} = v^N \left(\frac{\sqrt{2\pi mkT}}{h} \right)^{3N},$$

where $v = ABC$ is the volume of the gas, agreeing exactly with the classical value, so that equation of state and specific heat are not altered by using the quantum theory. At lower temperatures, where we cannot replace the summation by an integration, there will be discrepancies; the gas here is said to be "degenerate." At the same time other features enter the situation, different

sorts of statistics known as the Bose and Fermi statistics, which we shall discuss later in other connections. We shall not work out the case of degeneracy here, since practically one cannot reach such low temperatures without liquefying the gas, and since we shall meet in the next chapter some corresponding situations in solids, which are actually attained, and are of much more physical interest.

Problems

1. For neon, experimentally, $b = 20.6$ c.c. for a mol. Find the equivalent diameter r_0 of the atoms, regarded as rigid spheres. Compare this with the sum of the quantities $n^2/(Z - S)$ for the two atoms.

2. Using our approximate methods of dealing with Van der Waals' attraction, and using the value of r_0 from Prob. 1, compute the constant A for neon. Compare with the experimental value of 0.21×10^{12} absolute units (you cannot expect very good agreement).

3. Using the experimental values of b and A for neon, from Probs. 1 and 2, draw a graph for the second virial coefficient as function of temperature. At what temperature does the graph cross the T axis, and what does this mean physically?

4. Carry out the expansion of Van der Waals' equation in virial form, showing that the second virial coefficient is as we have found, and computing the third virial coefficient as well.

5. Using Van der Waals' equation, plot a number of isothermals (lines of constant T , p being plotted against v). Choose both low and high temperatures. Use the constants given in Probs. 1 and 2 for neon. Note that at low temperatures the isothermals have a maximum and minimum, while at high temperatures they do not. As is well known, this maximum and minimum are not really present, but the region in which they occur is that in which gas and liquid are in equilibrium and exist as a mixture.

6. The critical point is that point where the maximum and minimum of the isothermals of Van der Waals' equation coincide, or where the first derivative of pressure with respect to volume at constant T has a double root. Compute the critical pressure, temperature, and volume, for neon, using the constants given in Probs. 1 and 2.

7. Hydrogen gas is confined in a container 10 cm. on a side. Find the order of magnitude of the temperature at which it would become degenerate; that is, the temperature at which most of the molecules would be in the lowest quantum state.

8. Compute the internal energy and entropy of a perfect gas by the classical theory.

CHAPTER XXXVII

NUCLEAR VIBRATIONS IN MOLECULES AND SOLIDS

In the last chapter, we have seen that in addition to the equation of state of gases, there was another range of phenomena which we could treat satisfactorily: the phenomena resulting from nuclear vibrations in molecules, leading to the vibrational specific heat, and in solids, leading to the equation of state and specific heat. The mathematical methods used in dealing with them are similar, so that they can be profitably treated together. At the same time, the question of the stationary states of vibrating molecules is of interest in itself, and can be easily taken up.

We shall begin with the problem of a crystalline solid; the extension to a molecule, which after all is not very different from a fragment of such a solid, is not hard to make. Our problem is to find pressure as function of temperature and volume, and specific heat. Ordinarily the measurements of the equation of state of a solid take the form of measuring the compressibility and thermal expansion: we express volume as a function of pressure and temperature, and have

$$\begin{aligned}\text{compressibility} = \kappa &= -\frac{1}{v} \left(\frac{\partial v}{\partial p} \right)_T, \\ \text{thermal expansion} &= \frac{1}{v} \left(\frac{\partial v}{\partial T} \right)_p.\end{aligned}$$

We shall thus compute these quantities. Now a solid, unlike a gas, behaves in a perfectly normal way at the absolute zero of temperature. Its volume is finite and definitely determined, being given from the equilibrium positions of its own atoms and molecules, which all pack closely enough to be in their equilibrium positions, since they have no kinetic energy. If external pressure is applied, the volume will decrease, and we can compute the compressibility. Temperature will not greatly change these quantities: temperature agitation slightly increases the volume, and makes the crystal more compressible, but these effects are

small enough to be treated as perturbations of the state at absolute zero. Hence we begin by considering the crystal without temperature agitation.

270. The Crystal at Absolute Zero.—The energy of a crystal at the absolute zero, when its atoms are in perfectly definite positions, is simply the sum of the interaction energies for all pairs of atoms. In the position of equilibrium, this energy must be a minimum with respect to any possible small deformation of the crystal. Thus each separate atom is in equilibrium with respect to a slight displacement, keeping all other atoms fixed, so that it is at a minimum of potential, and could execute vibrations about this position of equilibrium, which to a first approximation would be simple harmonic. But there are other sorts of distortion to consider. For instance, we may decrease the whole volume slightly, moving the atoms closer together but preserving their relative arrangement, and the energy must be a minimum with respect to such a distortion. It is this which particularly interests us in computing compressibilities. Now in a very simple crystal lattice, if the volume is decreased, the atoms will still have just the same arrangement. Thus NaCl has a cubic lattice, Na and Cl ions being found alternately at the corners of cubes, and squeezing the whole lattice would merely decrease the size of the cubes. The same thing is true of the simpler metals. It is easy to see that it is not always the case; a crystal composed of molecules rather loosely tied together would, under compression, have the molecules forced closer together without much change in the dimensions of each molecule. We do not consider such complicated cases, however, but rather assume that all interatomic distances r_0 are proportional to the dimension δ of the crystal as a whole.

Let us assume, then, a cube of crystal, of side δ , a quantity which depends on the pressure. Let this cube contain N atoms. Now let us assume that the potential energy of the force between two atoms at distance r is the sum of two terms: an attractive term, negative in sign, proportional to $1/r$ for ionic crystals, or exponential in r for valence crystals; and a repulsive term, positive, and varying exponentially with r . The total energy of the crystal is the sum of all interatomic potential terms. This sum, for the exponential terms, is easy to compute. For these terms fall off so rapidly that practically only the nearest neighbors contribute appreciable terms to the energy. Thus

we simply take each of the N atoms, and sum up the exponential terms to its s nearest neighbors. This, as we readily see, gives each pair counted twice over, so that the sum is $\frac{1}{2}NsAe^{-ar}$, where r is the distance to the nearest neighbor, A and a are constants, or $\frac{1}{2}NsAe^{-\alpha\delta} = Ce^{-\alpha\delta}$, where $\alpha = ar/\delta$, since r is proportional to δ . For the inverse power attraction between ions, we cannot confine ourselves to nearest neighbors, since the forces fall off too slowly. Since each term in the energy is proportional to an inverse interatomic distance, and therefore to $1/\delta$, however, the energy will likewise be proportional to $1/\delta$, and the coefficient can be calculated by a proper method of summation over all ions.

Having the total energy, it is easy to compute the compressibility. We consider the ionic crystal, where the energy has the form

$$-\frac{K}{\delta} + Ce^{-\alpha\delta}. \quad (1)$$

We note that $dE = -pdv$, where E is the energy, p the pressure, $v = \delta^3$ is the volume, so that

$$p = -\frac{dE}{dv} = -\frac{dE}{d\delta} \frac{d\delta}{dv} = \left(-\frac{K}{\delta^2} + C\alpha e^{-\alpha\delta} \right) \frac{1}{3\delta^2}. \quad (2)$$

To compute the compressibility, we note that

$$\frac{dp}{dv} = \frac{dp}{d\delta} \frac{d\delta}{dv} = \frac{4K}{9\delta^7} - \frac{C(\alpha^2 + 2\alpha/\delta)e^{-\alpha\delta}}{9\delta^4},$$

from which we get the compressibility by the definition $\kappa = -\frac{1}{v} \frac{dv}{dp}$. Now we are interested in the properties of the solid at zero pressure. Setting the expression above for p equal to zero, a particular value of δ is determined, giving the volume at zero pressure. In turn, we substitute this into the compressibility, obtaining

$$\kappa_0 = \frac{-9\delta_0^4}{(2 - \alpha\delta_0)K}. \quad (3)$$

When we remember that $\alpha\delta_0 = ar_0$, and see from Chap. XXXV that this, for equilibrium, is about 8, we have approximately

$\kappa_0 = \frac{9\delta_0^4}{6K}$. This shows among other things that we may expect that in a series of similar crystals, as the alkali halides, where K may be expected to be the same for all, the crystals with larger grating spaces will be more compressible, since they will have larger δ_0 's for the same number of atoms.

271. Temperature Vibrations of a Crystal.—The atoms of a crystal will, of course, vibrate at temperatures above the absolute zero. The problem is very similar to that which we have already considered in Chaps. XI and XII, where we had particles coupled together, and considered their vibrations by means of normal coordinates. Here for sufficiently small amplitudes the potential energy will be quadratic in the displacements of all the particles, so that we can again introduce normal coordinates, though this will break down for too high temperature. We confine our discussion to the case where it can be done. There will be as many normal vibrations as there are coordinates; thus, with N atoms, each having 3 coordinates, there will be $3N$ normal coordinates. These coordinates can now execute simple harmonic vibrations, and the superposition of all these vibrations, each with its appropriate amplitude and phase, is the temperature agitation of the crystal. On classical mechanics, each of the normal coordinates can vibrate with any arbitrary amplitude, the actual magnitude depending in thermal equilibrium on the temperature. As a result of this, we have what is called equipartition of energy between the coordinates: each one, on the average, at temperature T , has kinetic energy $\frac{1}{2}kT$, potential energy equal to this, so that the total energy is kT , and the total thermal energy of the crystal $3NkT$, leading to a specific heat $3Nk$, which is given by Dulong and Petit's law, an experimental law for the specific heat of solids. We can easily prove this result. For example, to follow the method of the last chapter, we may compute the free energy:

$$e^{-\frac{F}{kT}} = \frac{1}{h^{3N}} \int \cdots \int e^{-\frac{H(q_1 \cdots p_{3N})}{kT}} dq_1 \cdots dp_{3N},$$

where the q 's are the normal coordinates, and the p 's are their conjugate momenta. Now from Chap. XIII we see that the

kinetic energy is equal to $T = \sum_{k=1}^{3N} \frac{\mu_k}{2} \dot{q}_k^2$, where μ_k is a mass coefficient.

cient connected with the k th normal coordinate, so that $p_k = \dot{q}_k/\mu_k$. Also, if ω_k is the angular frequency of this normal coordinate, $V = \sum \frac{\mu_k}{2} \omega_k^2 q_k^2$, so that $H = T + V$ is a sum of squares of momenta and coordinates. Thus we have

$$e^{-\frac{F}{kT}} = \frac{1}{h^{3N}} \int_{-\infty}^{\infty} e^{-\frac{p_1^2}{2\mu_1 kT}} dp_1 \cdots \int_{-\infty}^{\infty} e^{-\frac{\mu_{3N} \omega_{3N}^2 q_{3N}^2}{2kT}} dq_{3N}.$$

The limits of integration for the normal coordinates are given as $-\infty$ and ∞ . This of course is not correct; the whole method of normal coordinates becomes impossible if the q 's are even moderately large. But for low temperature the integrand will be small before the q 's have attained such large values, so that we can just as well integrate to infinity. We then have, performing the integration,

$$e^{-\frac{F}{kT}} = \frac{1}{h^{3N}} (\sqrt{2\pi\mu_1 kT}) \cdots (\sqrt{2\pi\mu_{3N} kT}) \left(\sqrt{\frac{2\pi kT}{\mu_1 \omega_1^2}} \right) \cdots \left(\sqrt{\frac{2\pi kT}{\mu_{3N} \omega_{3N}^2}} \right)$$

$$F = -kT \sum_{i=1}^{3N} \ln \frac{2\pi kT}{h\omega_i} = -3NkT \ln T + kT \sum_i \ln (h\nu_i/k), \quad (4)$$

where $\nu_i = \frac{\omega_i}{2\pi}$. From this, $C_v = -T \frac{\partial^2 F}{\partial T^2} = 3Nk$, as we have stated.

We may, however, treat the problem by the wave mechanics. Here we consider the normal coordinates to be the coordinates in Schrödinger's equation. On account of the fact that the Hamiltonian is a sum of terms, one connected with each normal coordinate, Schrödinger's equation is separable, and the equation for each coordinate is just like that for a linear oscillator. It will then have stationary states given by $E_{ki} = (k + \frac{1}{2})h\nu_i$, giving the k th energy level of the i th normal coordinate. The total energy is the sum of all the energies of the oscillators, or

$$\sum_{i=1}^{3N} (k_i + \frac{1}{2})h\nu_i. \quad \text{Using the quantum expression for free energy,}$$

we then have

$$\begin{aligned}
 e^{-\frac{F}{kT}} &= \sum_{k_1=0}^{\infty} \cdots \sum_{k_{3N}=0}^{\infty} e^{-\sum_{i=1}^{3N} (k_i + \frac{1}{2}) \frac{h\nu_i}{kT}} \\
 &= \sum_{k_1=0}^{\infty} e^{-(k_1 + \frac{1}{2}) \frac{h\nu_1}{kT}} \cdots \sum_{k_{3N}=0}^{\infty} e^{-(k_{3N} + \frac{1}{2}) \frac{h\nu_{3N}}{kT}}.
 \end{aligned}$$

These summations can be easily carried out: we have

$$\begin{aligned}
 \sum_{k=0}^{\infty} e^{-(k + \frac{1}{2}) \frac{h\nu}{kT}} &= e^{-\frac{1}{2} \frac{h\nu}{kT}} \left[1 + e^{-\frac{h\nu}{kT}} + \left(e^{-\frac{h\nu}{kT}} \right)^2 + \cdots \right] \\
 &= \frac{e^{-\frac{1}{2} \frac{h\nu}{kT}}}{1 - e^{-\frac{h\nu}{kT}}}.
 \end{aligned}$$

From this we have easily

$$\begin{aligned}
 F &= \sum_{i=1}^{3N} \frac{1}{2} h\nu_i + kT \sum_{i=1}^{3N} \ln \left(1 - e^{-\frac{h\nu_i}{kT}} \right), \\
 E &= \sum_{i=1}^{3N} \frac{1}{2} h\nu_i + \sum_{i=1}^{3N} \frac{h\nu_i}{e^{\frac{h\nu_i}{kT}} - 1}, \quad (5)
 \end{aligned}$$

and

$$C_v = k \sum_{i=1}^{3N} \frac{\left(\frac{h\nu_i}{kT} \right)^2 e^{\frac{h\nu_i}{kT}}}{\left(e^{\frac{h\nu_i}{kT}} - 1 \right)^2}.$$

Curves for E and C_v , its derivative, are shown in Fig. 76. The energy, for a single normal vibration, approaches the zero point energy $\frac{1}{2}h\nu_i$ at low temperature, where the exponential in the denominator of the second term is very large, so that the whole term is small. On the other hand, at high temperature, the exponential is small and can be expanded in power series, giving

$$\begin{aligned}
 E &= \frac{1}{2} h\nu_i + \frac{h\nu_i}{1 + \frac{h\nu_i}{kT} + \frac{1}{2} \left(\frac{h\nu_i}{kT} \right)^2 + \cdots - 1} \\
 &= \frac{1}{2} h\nu_i + kT \left(1 + \frac{1}{2} \frac{h\nu_i}{kT} + \cdots \right)^{-1} \\
 &= kT + \text{terms in } 1/T, 1/T^2, \dots \quad (6)
 \end{aligned}$$

for one normal vibration, or an energy of $3NkT$ at high temperatures for the whole crystal, leading to a specific heat which approaches zero at the absolute zero, but approaches the classical value $3Nk$ at high temperatures.

To get more information about the specific heat, we must consider the values of the natural frequencies ν_i . It is not impossible, with simple crystals, to solve the problem of the secular equation connected with the normal coordinates exactly, and find the ν 's. Born has done this, and has found the corresponding specific heat, in general agreement with experiment. But there are two simpler approximation methods which have been used, giving fairly good results. First, Einstein assumed

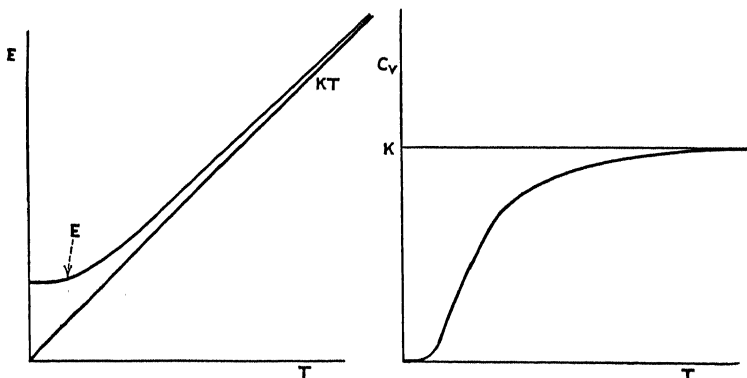


FIG. 76.— E and C_v for one degree of freedom, quantum theory of linear oscillator.

that all natural frequencies were the same, so that C_v was merely $3N$ times an expression like that above. This gives a specific heat which is zero at the absolute zero, but rises in the region of temperature given by $kT = h\nu$ to the value $3Nk$, given by the classical theory. Qualitatively, this is in accord with observations, but quantitatively at low temperatures it is not accurate. Obviously at low temperatures Einstein's formula gives $C_v = 3Nk \left(\frac{h\nu}{kT} \right)^2 e^{-\frac{h\nu}{kT}}$, whereas experimentally C_v is proportional to T^3 at low temperatures. Debye has given an explanation of this, which we shall sketch briefly. In the first place, as we have seen in Chap. XII, the frequencies of the lower overtones of a vibrating string agree with those of the equivalent weighted string. The same thing is true with a vibrating solid, the corresponding three-dimensional case. While the higher overtones

do not agree, still Debye finds that he does not introduce very serious errors by assuming that the frequencies of the actual vibrations are those of the continuous vibrating solid, or, by analogy with the membrane, by a formula of the nature of $\omega = \pi \sqrt{\frac{T}{\mu} \left(\frac{k^2}{X^2} + \frac{l^2}{Y^2} + \frac{m^2}{Z^2} \right)}$, where T is an elastic constant, μ a density, k, l, m integers, X, Y, Z the dimensions of the solid. This, being a continuous medium, has an infinite number of vibrations; but Debye assumed that the first $3N$ of these frequencies approximated the $3N$ normal coordinate frequencies. Now it is easy, as in Prob. 4, Chap. XV, to show that the number of overtones, or normal coordinates, with frequency between ω and $\omega + d\omega$ is proportional to $\omega^2 d\omega$, up to ω_{\max} , the frequency of the highest overtone, and after that it is zero. Hence the summation for C_v can be replaced by an integration,

$$\text{constant} \int_0^{\nu_{\max}} \frac{\left(\frac{h\nu}{kT} \right)^2 e^{\frac{h\nu}{kT}}}{\left(e^{\frac{h\nu}{kT}} - 1 \right)^2} d\nu,$$

leading to Debye's formula for the specific heat. At high temperature, it gives results about like Einstein's, with ν_{\max} substituted for ν . At low temperature, however, we have

$$C_v = \text{constant} \int_0^{\nu_{\max}} \frac{\nu^4}{T^2} e^{-\frac{h\nu}{kT}} d\nu.$$

For low temperatures, the exponential is very small, even for frequencies much below ν_{\max} , and we can without error integrate to infinity. Thus we have

$$\begin{aligned} C_v &= \text{constant} T^3 \int_0^{\infty} x^4 e^{-x} dx, \text{ where } x = \frac{h\nu}{kT} \\ &= \text{constant} T^3, \end{aligned}$$

giving correctly the behavior of the specific heat at low temperatures.

272. Equation of State of Solids.—From our free energy, we can find the equation of state, as well as the specific heat. We must note two things, however. First, the Hamiltonian we have used, whether on classical or quantum theory, is not the whole energy, but merely the part connected with thermal vibrations.

To this we must add $H_0(v)$, the energy when all atoms are stationary, which we have already computed in finding the compressibility at the absolute zero. Secondly, we must note that the frequencies ν_i of the overtone vibrations will depend on the volume. For as the crystal is compressed, and the atoms are forced closer together, they will be in regions of stronger force, and will vibrate with higher frequency. Remembering both these facts, we have on classical theory

$$\begin{aligned}
 F &= -3NkT \ln T + kT \sum_i \ln \frac{h\nu_i}{k} + H_0, \\
 p &= -\left(\frac{\partial F}{\partial v}\right)_T = -\frac{dH_0}{dv} - \sum_i \frac{kT}{\nu_i} \frac{\partial \nu_i}{\partial v}, \\
 &= p_0 - 3NkT \left(\frac{1}{\nu_i} \frac{\partial \nu_i}{\partial v} \right), \tag{7}
 \end{aligned}$$

where p_0 is the pressure at absolute zero, which we have already computed. The other term is the additional, thermal pressure, the average being taken over all overtones. Now experimentally it is found that the frequencies increase with decreasing volume, something like $\nu \propto 1/v^\alpha$, where α is an exponent of the order of magnitude of 2, so that $\frac{\partial \nu}{\partial v} = -\frac{\alpha \nu}{v}$, $\left(\frac{1}{\nu_i} \frac{\partial \nu_i}{\partial v} \right) = -\frac{\alpha}{v}$. We thus have approximately

$$p = p_0 + \frac{3\alpha NkT}{v}, \tag{8}$$

showing that the pressure is the sum of the pressure at zero temperature, arising from the atomic forces, and an additional term which is like the pressure of a perfect gas of the same number of atoms, only several times as great.

Using wave mechanics, we have

$$\begin{aligned}
 p &= p_0 - \frac{\partial}{\partial v} \sum_i \left[\frac{1}{2} h\nu_i + kT \ln \left(1 - e^{-\frac{h\nu_i}{kT}} \right) \right] \\
 &= \left(p_0 - \sum_i \frac{1}{2} h \frac{\partial \nu_i}{\partial v} \right) - \sum_i \frac{h \frac{\partial \nu_i}{\partial v}}{e^{\frac{h\nu_i}{kT}} - 1}. \tag{9}
 \end{aligned}$$

The pressure at absolute zero of temperature has a small correction, on account of the terms $\frac{1}{2}h\nu_i$. And the thermal term in the

pressure has quite a different form from what it has in classical theory. At high temperature, this approaches the classical value, but at low temperatures it goes down to zero, the thermal pressure being less than classically. This is a real case of degeneracy, met in practice, unlike the degeneracy of gases, which we met in the preceding chapter, and which cannot be actually realized.

273. Vibrations of Molecules.—The small vibrations of a molecule behave just as do those of a crystal, the only difference being that the number of degrees of freedom, and hence of normal coordinates, is small. Molecules have vibrational specific heat, falling off to zero at low temperatures as with crystals. In fact, the vibrational frequencies are generally such that at room temperature the oscillations are not excited, and the vibrational specific heat is zero, becoming apparent only at rather high temperatures. The particular interest in the vibrations arises from band spectra. A molecule can jump from one vibrational state to another, emitting radiation, generally in the infra-red, whose frequency is given by the difference of the vibrational energies. Or a molecule in an excited electronic state can jump to another electronic state, with simultaneous change of vibrational quantum number, emitting light in the visible or ultra-violet. Both types of spectra show the appearance of bands, from which they are named. A general discussion of band spectra is beyond the scope of this book, but we can at least consider a little the type of vibrational levels to be expected, and draw some conclusions about the sort of spectra we expect.

We have already seen that there will be a number of normal coordinates, each acting like a single linear oscillator, and therefore having energy levels $(k_i + \frac{1}{2})h\nu_i$, so that the whole energy will be a sum over the normal coordinates of such expressions. With N atoms, we might suppose that there would be $3N$ such vibrations. As a matter of fact, however, this is not the case: three out of the $3N$ coordinates are taken up in describing the position of the center of gravity, and either two or three in describing the orientation of the molecule, two for a linear molecule, as a diatomic one, and three, as for instance Euler's angles, for a nonlinear, polyatomic molecule, so that only $3N - 5$ or $3N - 6$ are left for vibration. Thus a diatomic molecule ($N = 2$, $3N - 5 = 1$) has only one mode of vibration, that

in which both its atoms simultaneously move toward, or away from, each other; a nonlinear triatomic molecule has three; and so on. We could neglect these corrections with crystals without serious error, on account of the enormous size of N . We shall commence our discussion with the diatomic molecule, and then shall indicate briefly the nature of the extension to the polyatomic case.

274. Diatomic Molecules.—We have already seen, in Fig. 75, the form of the potential curve of a diatomic molecule, as function of the distance of separation of the atoms, the one coordinate with respect to which we have vibration. The quantity $(r - r_0)$ plays the part of a normal coordinate directly, so that to the extent to which we can use normal coordinates at all, we must replace the actual potential curve by a parabola, approximating it as well as possible at the minimum. The vibrational levels would then be $(n + \frac{1}{2})h\nu$, where ν is the vibrational frequency in the parabolic potential. We can go farther in the solution in this particular case, solving the vibrational problem exactly if the potential is given by a curve of Morse's type, though we shall not do that here. It is necessary, however, to consider rotation as well as vibration, and we shall carry this discussion through.

Let us start with the general case of two atoms of masses m_1, m_2 , with a potential $V(r)$ between them, where r is the distance between. The Hamiltonian is $(p_1^2/2m_1) + (p_2^2/2m_2) + V(r)$. Let us, however, separate off the coordinates of the center of gravity. Let $x_1y_1z_1$ be the coordinates of the first particle, $x_2y_2z_2$ of the second. Then we introduce new coordinates,

$$\begin{aligned} X &= \frac{m_1x_1 + m_2x_2}{m_1 + m_2}, \text{ etc.,} \\ \xi &= x_2 - x_1, \text{ etc.} \end{aligned} \quad (10)$$

Here X, Y, Z are the coordinates of the center of gravity, ξ, η, ζ the relative coordinates, so that $\xi^2 + \eta^2 + \zeta^2 = r^2$. In terms of these coordinates we readily find that

$$T = \frac{1}{2}(m_1 + m_2)(\dot{X}^2 + \dot{Y}^2 + \dot{Z}^2) + \frac{1}{2}\mu(\dot{\xi}^2 + \dot{\eta}^2 + \dot{\zeta}^2),$$

where $\mu = \frac{m_1m_2}{m_1 + m_2}$. Thus the Hamiltonian is

$$\frac{p_x^2 + p_y^2 + p_z^2}{2(m_1 + m_2)} + \frac{p_\xi^2 + p_\eta^2 + p_\zeta^2}{2\mu} + V(\xi\eta\zeta). \quad (11)$$

This is already separated, on either classical or quantum theory, the first term giving the translational energy of the molecule of mass $(m_1 + m_2)$, the second being like the motion of a particle of mass μ in a central field $V(r)$. This latter problem can be solved in wave mechanics by introducing polar coordinates, just as we did with the general central field problem. As in Chap. XXXIII, the solution is

$$u = e^{\pm im\phi} P_l^m(\cos \theta) R(r),$$

where $y = rR$ satisfies the equation

$$\frac{\hbar^2}{8\pi^2\mu} \frac{d^2y}{dr^2} + \left[E - V(r) - \frac{l(l+1)\hbar^2}{8\pi^2\mu r^2} \right] y = 0, \quad (12)$$

l being an integer. Qualitatively, however, the situation is quite different as far as the function of r is concerned. With the electrons in atomic structure, we had a quantum number for this equation representing the number of radial nodes, which we might call n_r , and then had $n_r + l + 1 = n$, the total quantum number. We found that the energy, in the hydrogen case, was proportional to $1/n^2$. That is, terms of the same n_r had very different energy, depending on the azimuthal quantum number. Here, however, on account of the fact that μ is much larger than the mass of the electron, the term $\frac{l(l+1)\hbar^2}{8\pi^2\mu r^2}$ is very much smaller than in the problem of atomic structure. As a result, the energy depends only slightly on l , determining the rotation. The number n_r plays the part of the vibrational quantum number, and we have approximately $E = (n_r + \frac{1}{2})h\nu$, but with a small correction term. We can get a first approximation to the correction very easily, from perturbation methods. We first solve without the term depending on l , then introduce it as a perturbation in the energy. As we have seen, the perturbed energy is, to the first approximation, the unperturbed energy plus the mean value of the perturbed energy over the unperturbed wave function. Thus the rotational contribution to the energy is $\frac{l(l+1)\hbar^2}{8\pi^2\mu r^2}$. If we note that the amplitude of vibration is not very large, so that r does not vary a great deal, we see that $\frac{1}{r^2}$ is approximately equal to $\frac{1}{r_0^2}$, where r_0 is

the equilibrium distance, so that, setting $I = \mu r_0^2$, the moment of inertia, we have

$$E = \left(n_r + \frac{1}{2}\right)h\nu + \frac{l(l+1)h^2}{8\pi^2 I}. \quad (13)$$

This gives a set of vibrational levels (given by n_r), each broken up into a group of rotational levels (given by l), which are fairly closely spaced.

Electronic band spectra are emitted when we have transitions between two such sets of levels. To the energy above we must add the electronic energy; that is to say, we have counted our energy so far from the minimum of the potential curve, and this is different for different states. Two such electronic energy levels, with the corresponding vibrational and rotational levels indicated, are plotted in Fig. 77. In considering transitions, there is a selection principle which must be considered, as with atomic systems, l being able to change only by ± 1 or 0 units. As a result, bands have three branches, called P , Q , R , corresponding to the three possible changes of l (P corresponding to decrease of l by one unit, Q no change, R increase). These branches are indicated in Fig. 77. It should be stated that actual band spectra are ordinarily much more complicated than this, on account of a number of factors which we have not yet considered, as multiplet structure and the existence of electron spin.

275. Specific Heat of Diatomic Molecules.—Using an approximate energy expression for a diatomic molecule, we can find the specific heat of a gas of such molecules. The translational part of the energy separates from the rotational and vibrational part, as we have seen, giving a contribution $\frac{3}{2}Nk$ to the specific heat. The whole free energy is given by

$$e^{-\frac{F}{kT}} = \left(\frac{\sqrt{2\pi mkT}}{h}\right)^{3N} v^N \sum_{n's, l's} e^{-\sum_i \left(\left(n_i + \frac{1}{2}\right)h\nu_i + \frac{l_i(l_i+1)h^2}{8\pi^2 I_i} \right) / kT} \quad (14)$$

The summation over i is over all molecules, so that each vibrational quantum number has a factor $\sum_{n=0}^{\infty} e^{-\frac{(n+1/2)h\nu}{kT}}$, and each

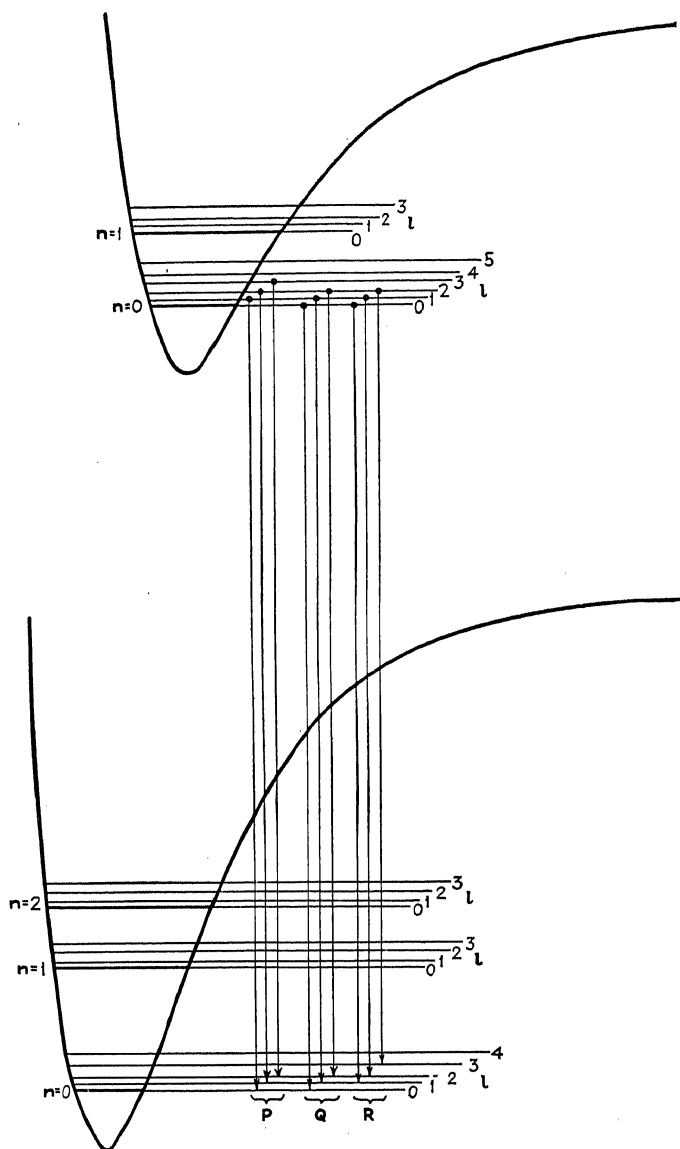


FIG. 77.—Energy levels and transitions for diatomic molecule.

rotational one a term $\sum_{l=0}^{\infty} e^{-\frac{l(l+1)\hbar^2}{8\pi^2 I k T}}$. The first we have already

considered, and have seen that its contribution to the specific heat is appreciable only at high temperatures. The last term, arising from the rotation, can be easily calculated at high temperatures, where the rotational energy is large, and we must consider states of large l 's. We must first notice that for each value of l there are $(2l + 1)$ substates, corresponding to different m 's, all of the same energy, so that each term must actually be counted $(2l + 1)$ times in the summation. Then we replace the summation by an integration, and neglect unity compared with l , finding

$$\int_0^{\infty} 2l e^{-\frac{l^2 \hbar^2}{8\pi^2 I k T}} dl = \frac{8\pi^2 I k T}{\hbar^2},$$

so that each molecule contributes to the free energy a term $-kT \ln \left(\frac{8\pi^2 I k T}{\hbar^2} \right)$ on account of its rotation, and to the specific heat the amount k , just the amount corresponding to two degrees of freedom according to the equipartition of kinetic energy. At low temperatures, we may not replace the sum by an integral, and the rotational specific heat proves when worked out to decrease to zero at the absolute zero, just as the vibrational specific heat does. This is not ordinarily observed for any gas except hydrogen, for it comes at so low a temperature that the gas condenses, but in hydrogen the phenomenon occurs at room temperature, on account of the small moment of inertia, correspondingly wide spacing of the rotational levels, and high temperature necessary to acquire an energy comparable with this spacing.

276. Polyatomic Molecules.—The theory of polyatomic molecules differs from that for diatomic molecules in several ways, though we shall not go into the question in detail. First, there are several different fundamental frequencies of vibration, corresponding to the various normal coordinates, so that the system of energy levels and the band spectrum are more complicated. Nevertheless, in some fairly simple cases it is possible to analyze the band systems, identifying the empirical frequencies with the corresponding modes of vibration. Next, the rotational

levels are more complicated, on account of the lack of symmetry of the problem. We can no longer solve the rotational problem immediately in polar coordinates. We could do it before because the rotating molecule would act like a symmetrical top, symmetrical about the internuclear axis, and the energy levels depended only on the moment of inertia about axes at right angles to this axis. But with a polyatomic molecule, unless it happens to be linear, this simplification is not present. The problem resembles a nonsymmetrical top, and as we saw in discussing the motion of a rigid body in classical mechanics, this introduces great complications. It can be approximately solved, but we shall not do it. One fact, however, can be stated at once about the rotational levels: their spacing depends in general on the reciprocal of the moment of inertia, and as we go to more and more complicated molecules, with larger and larger moment of inertia, the rotational levels are more and more closely spaced, so that the classical theory becomes more and more accurate. This is unfortunate for band spectroscopy, since the rotational lines in the spectrum are so closely spaced, in even moderately heavy molecules, that they cannot be resolved spectroscopically, the band appearing continuous rather than full of discrete lines. As a final point in connection with non-linear polyatomic molecules, the specific heat connected with rotation proves to be $\frac{3}{2}k$ for each molecule, corresponding to $k/2$ for each of the three coordinates concerned in determining the orientation of the molecule in space.

Problems

1. Given a crystal whose energy at the absolute zero is $-K/\delta + b/\delta^n$, where δ is a linear dimension, b and n are constants, find the compressibility at zero pressure, as function of K , δ_0 , and n .

2. In rock salt, ions of Na^+ and Cl^- are arranged alternately on a cubic lattice, the equilibrium distance between successive Na^+ ions being 5.7×10^{-8} cm. If δ is chosen as the distance between successive Na^+ ions, so that $\delta_0 = 5.7 \times 10^{-8}$, the energy in the cube of volume δ^3 may be approximated $-K/\delta + b/\delta^n$, where $K = 13.94 e^2$, if e is the charge on an electron, and n is about 9 units. Find the compressibility at the absolute zero, in reciprocal dynes per square centimeter, comparing with the observed value (between 3 and 4×10^{-12}).

3. Using the figures of Prob. 2, compute the energy required to break up 1 gm. mol. of NaCl into Na^+ and Cl^- ions (that is, to make δ infinite). Compare with the experimental value, in the neighborhood of 180 kg. cal. per gram-molecule.

4. Using Dulong and Petit's law, compute the specific heat at ordinary temperatures of copper and lead, and compare with the experimental values.

5. Compute and plot Einstein's specific heat curve, assuming $h\nu/k = 100^\circ$ abs.

6. Prove that Einstein's specific heat curve approaches the classical value at high temperature.

7. Using the expression $p = p_0 + (3\alpha NkT/v)$ for the pressure of a solid at any temperature in terms of its pressure at the absolute zero, find the thermal expansion, showing that it is approximately equal to $\alpha C_v \kappa_0/v$, where κ_0 is the compressibility. (Suggestion: Set the pressure equal to zero, find the volume as function of temperature, expressing p_0 as a function of volume by means of the compressibility.)

8. By methods like those of Prob. 7, show that the fractional change of compressibility with temperature, $\frac{1}{\kappa_0} \frac{\partial \kappa}{\partial T}$, is to our approximation equal to the thermal expansion. From the figures given above for NaCl, compute the order of magnitude of these quantities for this crystal.

9. Show that the thermal expansion is given by $\alpha C_v \kappa_0/v$ even at low temperatures where C_v must be given by the quantum theory rather than the classical formula.

10. Solve the problem of the vibrational levels of a molecule whose potential energy is $De^{-2au} - 2De^{-au}$, where $u = r - r_0$, without rotation. If R is r times the radial wave function, and $y = e^{-au}$, set up the differential equation for R , using y as independent variable, showing that it is

$$\frac{1}{y} \frac{d}{dy} \left(y \frac{dR}{dy} \right) + \frac{8\pi^2\mu}{a^2 h^2} \left(\frac{E}{y^2} + \frac{2D}{y} - D \right) R = 0.$$

Treat this equation like Schrödinger's equation for the hydrogen atom, letting $R = e^{-dy} (2dy)^{1/2} F(y)$, where $d = \frac{2\pi\sqrt{2\mu D}}{ah}$, $E = -\frac{a^2 h^2 b^2}{32\pi^2\mu}$. Obtain the differential equation for F , showing that the series solution breaks off to give a finite polynomial if the energy is given by

$$E_n = -D + \left(n + \frac{1}{2} \right) h\nu - \left(\frac{h^2 \nu^2}{4D} \right) \left(n + \frac{1}{2} \right)^2,$$

where ν is the frequency of classical vibration about r_0 , equal to $\frac{a}{2\pi} \sqrt{\frac{2D}{\mu}}$.

CHAPTER XXXVIII

COLLISIONS AND CHEMICAL REACTIONS

We have been considering the interactions of atoms in molecules, gases, and solids, under the action of their interatomic forces. There are a few special interesting and unusual cases of interaction, which we shall consider in the present chapter. Our discussion will be mostly qualitative, since an exact treatment is very difficult and complicated. We can classify the problems we are taking up by remembering that there is a potential energy function acting between atoms, depending on the electronic state. It is this energy which is responsible for the formation of molecules, and for their mutual attractions and repulsions, as we have seen. But there is one feature we have not considered: each different electronic state has a different interatomic energy connected with it. Thus an excited atom is much larger than a normal one, so that two such atoms begin to exert valence and repulsive forces on each other at much larger distances than atoms in their normal states. For any given problem, we have an infinite set of electronic energy levels as functions of the nuclear coordinates, as we have seen for the diatomic molecule in Fig. 77. Now all the problems we have spoken of (except for the emission of electronic bands, discussed in Fig. 77), have been problems in which there was no change of electronic quantum numbers. Ordinarily the electrons have been in their lowest stationary states, and the corresponding potential energy is the one used in discussing molecular formation and interatomic forces. We now classify the problems we shall take up in the present chapter into two groups: (1) those in which the electronic quantum numbers do not change; (2) those in which the electronic quantum numbers do change, with of course a compensating change in something else. Many chemical reactions are examples of (1), and we consider them first.

277. Chemical Reactions.—Probably the simplest chemical reactions to discuss are bimolecular gas reactions, in which

two molecules collide, are transformed into one or more other molecules, and the resulting molecules separate. The simplest case of two single atoms colliding does not ordinarily lead to the formation of a molecule at all, particularly if the atoms are not excited. For then there is a potential energy curve between them like Fig. 75. Considering the atoms as moving according to classical mechanics in this potential field, we see that if they start toward each other from infinity with finite kinetic energy, they will approach to a "perihelion" distance, separate, and finally go to infinity again, with the same kinetic energy as before. The only way for them to become bound would be for them to lose vibrational kinetic energy when close together, at a distance approximately r_0 , so that they would begin vibrating about the distance r_0 . This could conceivably happen if they radiated while close together, but calculation shows this to be most improbable. Actually the only important mechanism by which such binding occurs is a collision by a third particle, part of the kinetic energy being imparted to this particle, resulting in the binding of the atoms to form a molecule. Thus this reaction is not of the simple type we wish to consider.

One can really get bimolecular reactions, however, if at least one of the colliding molecules is diatomic. For example, it would be possible for HBr colliding with H to change into H_2 and Br, or *vice versa*. And such reactions can occur without change of electronic quantum number, the electrons always staying in their lowest state. To understand them, we need merely consider the potential energy function as it depends on the coordinates of all three atoms. In discussing the motion of systems of particles, we have seen that it is helpful to think of the potential energy as plotted in a many-dimensional space, and to imagine a ball rolling over this many-dimensional energy surface. In this case, we have the nine coordinates of our three particles, and the energy is to be plotted as a function of these nine variables. While it is impossible to visualize the whole diagram, we can easily describe some features of it. First, there will be regions of space corresponding to one H atom near the Br atom, but with the other H atom far off. Here the potential will depend only on the relative distance of the adjacent H and Br, the potential as function of the distance looking like the familiar energy of a diatomic molecule, and leading to a low total energy of the system if they are at the

proper distance r_0 . If the atoms move either closer together or farther apart the energy will increase. But there is also an entirely different region of the nine-dimensional space in which the two H's are near together, and the Br is at a great distance. This again will give a low potential energy, since the H's can be bound together at a suitable distance to form a molecule H_2 . The larger part of the coordinate space, however, will correspond to all three atoms being separated from each other, and will have a constant potential energy, a plateau, with the two valleys we have spoken of. The two valleys connect with each other, in the region of space where all three atoms are near together, but calculation shows that to get from the bottom of one valley to that of the other, one must go over a considerable elevation, though not so high as the plateau. Now suppose we start with HBr and H, the HBr molecule having no vibrational energy, and the mutual kinetic energy of HBr and H being small. The corresponding rolling ball will be in the bottom of the HBr valley, rolling slowly along the valley (this corresponds to relative velocity of the two molecules, without vibration of the HBr). The direction of motion is toward the junction with the other valley, corresponding to $H_2 + Br$. If the ball is going fast enough, it will be able to rise over the pass separating the two valleys, and roll down the other side, resulting in the formation of H_2 , or in the reaction we are interested in. We see, in other words, that a necessary condition for the possibility of such a reaction is a potential energy with two separate valleys, so that the point representing the system can start in one and end up in the other. Of course, complicated reactions could have several valleys, but the principle is the same. And in every case it turns out to be true that there is an elevated pass between valleys. This means that molecules must have a certain kinetic energy before they can react. This is called activation energy, and it is ordinarily large enough so that only a few of the fastest molecules can react. Now Maxwell's distribution of velocity is such that the number of molecules with particularly high speed increases very rapidly with temperature. Thus, if the necessary kinetic energy is E , the chance of finding a pair of molecules with the required energy will depend on a quantity like $e^{-\left(\frac{E}{kT}\right)}$, a very small quantity, but increasing rapidly with temperature. It is evident, then, that the reaction velocity

will increase rapidly with temperature. We can estimate this by assuming the velocity V to be proportional to $e^{-\left(\frac{E}{kT}\right)}$. Then we at once see by differentiating that $(d \ln V)/dt = E/kT^2$, the equation of the so-called reaction isochore. In ordinary reactions at room temperature, E/kT is of the order of 20, showing that only very fast molecules can react. Then, setting $T = 300^\circ \text{ Abs.}$, the result is about $(d \ln V)/dt = 0.07$, showing that $\ln V$ increases by about 0.7 for 10° rise of temperature. Since 0.7 is approximately $\ln 2$, this means that reaction rates approximately double for 10° rise of temperature.

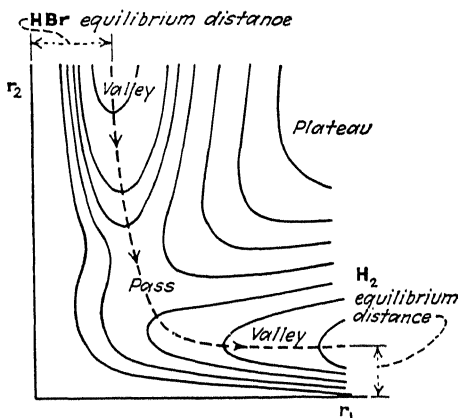


FIG. 78.—Potential energy of H-H-Br as function of H-Br distance (r_1) and H-H distance (r_2) for all three atoms in line. Dotted line indicates reaction $\text{HBr} + \text{H} \rightarrow \text{H}_2 + \text{Br}$, the height of the pass above the first valley being the activation energy. (Curve from H. Eyring and M. Polanyi, by permission.)

While the potential energy cannot be plotted in the general case, we can do so in a special case. Let all three atoms be in a straight line, in the order H-H-Br, and let the H-H distance be r_2 , the H-Br distance r_1 . Then we may plot energy as function of r_2 and r_1 , indicating it by contour lines. The result is a diagram as in Fig. 78, where the dotted line shows the path of the rolling ball in the reaction we have described.

278. Collisions with Electronic Excitation.—The chemical reaction we have just considered was of the type in which the electrons did not change from one stationary state to another. We can easily change the problem, however, to involve electronic transitions. Thus let us take even our simple case of two atoms colliding, but now with rather large kinetic energy. If this

energy is large enough, one or both of the atoms can become excited, the atoms separating with loss of enough kinetic energy to compensate for the increased electronic energy. This plainly

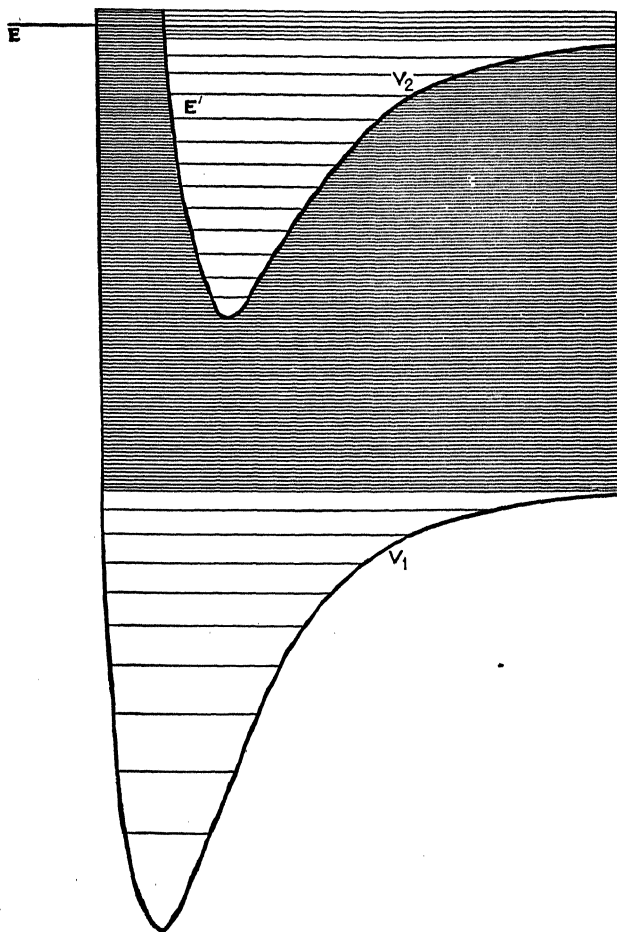


Fig. 79.—Potential energy curves for collision. The continuously shaded regions represent continuous distributions of energy levels, superposed on the discrete levels.

demands an initial kinetic energy greater than the lowest excitation potential of one or the other atom, a situation which is not ordinarily met in chemical problems, so that this does not represent a conventional chemical reaction, but which is easily realized in discharge tubes. To understand how the process can occur, we refer to Fig. 79. There we plot two electronic potential

energy curves, V_1 and V_2 , corresponding to different electronic energy, and evidently going to different energy levels of the atoms at infinite separation. The total energy is now taken to be E , greater than the energy V_2 of the excited electronic state. Now, of course, in the potential curve V_1 we have vibrational quantized levels of energy less than the value of V_1 at infinite r , but we also have continuous levels for higher energy, corresponding to the classical motions extending to infinity, and it is such a level which we consider. Similarly the potential curve V_2 has discrete and continuous levels, and in particular a continuous level corresponding to this same energy E . Now in Chap. XXXII we have seen that when there are two states of this sort, both corresponding to the same energy, there is a certain probability of passing from one to the other. Thus, proceeding by the perturbation method of variation of constants, we can start with the system in electronic state V_1 , but as time goes on the wave function will commence to take on some of the properties of electronic state V_2 . In other words, there is a certain chance that during the time of collision this transition will have taken place, and that on separation of the atoms they will be in the excited level corresponding to V_2 , with correspondingly small kinetic energy. As in Chap. XXXII, the rate of passing from one state to the other will depend on the non-diagonal matrix component of the energy between the two states; we shall shortly see how to compute this component.

Several other types of process similar to the one we have just considered are of importance. First there is the inverse process: excited atoms, corresponding to the potential curve V_2 , approach each other, lose their excitation energy, and separate with greater kinetic energy than they had before. This is called a collision of the second kind. Other more complicated collisions of the second kind are those in which most of the excitation energy of one atom passes in the collision to excitation energy of the other, only a relatively small amount being left for the change in kinetic energy. Such collisions prove to be much more likely than the ones with large change in kinetic energy. The essential reason is that if the kinetic energy changes greatly, the wave functions corresponding to the nuclear motion in the two states will have very different wave length. Then when we multiply the wave functions and integrate in finding the nondiagonal energy matrix component which produces the transition, we shall find that the

two waves of different wave length will have a product which averages practically to zero, producing a small integral. When the change of wave length (or of kinetic energy) is small, however, the two waves can interfere constructively, producing a large integral, and large probability of transition.

Still another sort of process is that involved in predissociation. Here one has a molecule in state V_2 , but with a discrete energy E' , so that it is a stable molecule, rather than two colliding atoms. This energy E' , however, corresponds at the same time to a continuous level of the potential function V_1 . There will then be a certain probability that the molecule will change over to this continuous level. If it does, it will at once dissociate, resulting in two atoms, with the potential curve V_1 , and large kinetic energy. Thus the molecular state E' is inherently unstable, as is a radioactive nucleus. At the same time, on account of the finite life time of the state, on account of the probability of dissociation, the energy levels will be broadened, as we saw in Chap. XXXII. This really means that a finite range of the continuous distribution of energy levels of the problem in the neighborhood of E' have properties suggesting the molecule and the state V_2 , rather than the colliding atoms and state V_1 . This broadening of energy levels is observed in the spectra of predissociating molecules.

An atomic phenomenon similar to predissociation is found when an atom has two excited electrons. In this case it is easily possible for it to have greater energy than the normal state of the ionized atom. Thus the discrete level we are considering lies at the same height as the continuous levels above the series limit of the ordinary spectrum, corresponding to an ionized atom in its lowest state, with an electron flying off with a positive kinetic energy. The system in the discrete level then has a certain probability of changing over to the continuous state, or, in other words, one of the electrons has a certain probability of escaping from the atom, the necessary energy being provided by the other excited electron falling down to a lower state. This spontaneous ionization of a doubly excited atom is called the Auger effect, and is the explanation of the fact that not many energy levels of doubly excited atoms above the series limit are observed. The levels are unstable, and have short lifetimes.

279. Electronic and Nuclear Energy in Metals.—We have not so far considered metals in detail. They present, however, a

feature so similar to what we have just discussed that it will pay to bring it up here. The characteristic feature of metals is that they contain free electrons, which can be accelerated under the action of an impressed electric field. Thus the wave function of a free electron of energy E , momentum components p_x, p_y, p_z , can be written

$$\psi = e^{\frac{2\pi i}{h}(p_x x + p_y y + p_z z - Et)}$$

If, however, there is a constant force F acting on the electron along the x direction, we should expect by classical mechanics that p_x would change with time, such that its time rate of change was F , or $p_x = p_{x0} + Ft$, while p_y and p_z would remain constant, E changing so that at all times it would be $\frac{(p_x^2 + p_y^2 + p_z^2)}{2m}$.

If then we set up the wave function

$$\psi = e^{\frac{2\pi i}{h}\left\{(p_{x0} + Ft)x + p_y y + p_z z - \int \left[\frac{(p_{x0} + Ft)^2 + p_y^2 + p_z^2}{2m}\right] dt\right\}}, \quad (1)$$

we readily find on carrying out the differentiations that it satisfies the differential equation

$$\left(-\frac{\hbar^2}{8\pi^2 m} \nabla^2 - Fx\right)\psi = -\frac{\hbar}{2\pi i} \frac{\partial \psi}{\partial t}, \quad (2)$$

or Schrödinger's equation corresponding to the potential energy $-Fx$, or the force F . This solution obviously does not correspond to a constant energy, but rather to a continuously increasing energy, as if we were passing continuously from one real stationary state of a free electron in the absence of field to another. In other words, the external field, regarded as a perturbation on the electron in free space, produces continual transitions, amounting to a uniform acceleration.

Now if there were nothing to stop it, this process of acceleration of the free electrons would go on for ever, their velocities and energies increasing without limit. The counteracting influence is the collisions which the electrons make with the atoms of the lattice, setting them into vibration, and losing energy themselves. These two processes, of acceleration of the electrons and collision, result in an equilibrium, giving a uniform drift velocity, and a uniform current as a result, and at the same time resulting

in a continuous process of increasing the lattice vibrations, which is simply the heating of the metal on account of its resistance. Investigation shows that the probability of collision with loss of the kinetic energy is proportional to the velocity of the electron. This results in an apparent resisting force proportional to the velocity, and the electrons act like particles subject to an external force field, and a resisting force. As we have seen in Chap. II, the resulting motion is a uniform drift with velocity (and hence current) proportional to the external field. This is the origin of Ohm's law.

Now the process of collision of the electrons with the ions of the lattice is essentially similar to the other collision processes we have been considering in the present chapter. Let us forget for a moment the accelerating action of the field, and simply assume that it has already raised the electrons to states of large momentum and kinetic energy, but that it is momentarily not acting. The crystal as a whole is now in a state of high electronic energy, but of low nuclear vibrational energy. There are, however, other states of the crystal of the same energy, corresponding to smaller electronic energy, but higher nuclear energy. In particular, these other states mostly correspond to no net momentum of all the electrons, or no electric current. There will be a probability of interaction between the original state and these other states, and in the course of time the system will pass over completely to these other states, the current being lost, its momentum dissipated, its energy converted into nuclear vibrational energy. Of course, some of the kinetic energy will remain with the electrons, since they have finite though small specific heat, and the system tends to thermal equilibrium. But the momentum, and the current, will be lost. In this case, unlike the case of collisions, the inverse process, in which the nuclei lose energy and change it into electronic energy, might happen enough to keep equilibrium, but it would never restore the current of its own accord. The reason is that this is such an excessively unlikely process, since of all the electronic states corresponding to the same energy, only one will correspond to having the electrons all moving in the same direction and cooperating to form a net current, while in most of them the electrons will be moving in all directions, and their currents will cancel. This process of dissipation of the current, an essentially irreversible process, is similar to the process of radiation of energy, which we have con-

sidered in Chap. XXXII, and in both cases the irreversibility arose in having a simple state change into a very complicated one. Such a change is always accompanied by an increase of entropy, and is a typical irreversible process of thermodynamics.

280. Perturbation Method for Interaction of Nuclei.—So far, we have not really applied quantum mechanics at all to our problem of nuclear interaction. We have simply assumed that potential energy functions of the positions of the nuclei can be set up, and that these are to be used for discussing the nuclear motions. In phenomena of our first type, including all our equation of state problems of the preceding chapters, and simple chemical reactions as well, the nuclei moved as if only one potential function had to be considered. On the other hand, we have just been discussing phenomena of a second type, in which there was appreciable probability of transition from one electronic level to another. We shall now analyze this problem more in detail, showing how the electronic and nuclear motions can be separated, to a first approximation exactly (leading to the phenomena of the first type), but to a higher approximation not quite perfectly (leading to the transitions of the second type). The reason why this very fortunate separation is almost exactly possible lies essentially in the great difference of mass between nuclei and electrons, resulting in differences of velocity and wave length. The electrons move so fast that they go through their orbits many times while the nuclei move a small distance, so that we may approximately solve for the motion of the electrons, assuming the nuclei to be fixed. In this solution, the coordinates of the nuclei will appear as parameters in the wave function. Similarly the energy levels of the electronic system will depend on the position of the nuclei. And it turns out to be actually true, as we have been assuming, that the electronic energy, as function of the position of the nuclei, plays the part of a potential for the nuclear motion.

It is not hard to show how the ideas we have just described lead to a separation of variables. Let the electronic coordinates be symbolized by x_i , the nuclear ones by X_j , the mass of an electron being m , and of the j th nucleus M_j . Then Schrödinger's equation may be written

$$\left(\sum_i -\frac{\hbar^2}{8\pi^2 m} \frac{\partial^2}{\partial x_i^2} + \sum_j -\frac{\hbar^2}{8\pi^2 M_j} \frac{\partial^2}{\partial X_j^2} + V(x_i, X_j) \right) \psi = -\frac{\hbar}{2\pi i} \frac{\partial \psi}{\partial t}$$

Now assume that $\psi = e^{-\frac{2\pi i}{h}Et} u(x_i, X_j) v(X_j)$, where $u(x_i, X_j)$ is the wave function for the electrons, assuming the nuclei fixed, and containing therefore the X_j 's as parameters, and $v(X_j)$ is the wave function for the nuclei, under the action of the potential arising from the electrons. That is, u is a solution of the equation

$$\left[\sum_i -\frac{\hbar^2}{8\pi^2 m} \frac{\partial^2}{\partial x_i^2} + V(x_i, X_j) \right] u(x_i, X_j) = \epsilon(X_j) u(x_i, X_j). \quad (3)$$

The Hamiltonian on the left is what we get by assuming the masses of the nuclei are infinite, so that they stay at rest. Since the potential function V depends on the X_j 's as parameters, the energy ϵ must also show this dependence. Now we take $\epsilon(X_j)$ as the potential for the nuclei; that is, v satisfies the equation

$$\left[\sum_j -\frac{\hbar^2}{8\pi^2 M_j} \frac{\partial^2}{\partial X_j^2} + \epsilon(X_j) \right] v(X_j) = E v(X_j), \quad (4)$$

where E is the whole energy of the system. For example, with a diatomic molecule, $\epsilon(X_j)$, the electronic energy as function of the nuclear position, is the potential curve we have often used, and E is the vibrational energy level, measuring actually the total energy, electronic and vibrational, and staying constant during the motion, the electronic energy decreasing when vibrational energy increases, and *vice versa*.

Let us now see what differential equation uv satisfies; it proves to be this product which approximates a wave function for the whole system. We have easily

$$\begin{aligned} & \left[\sum_i -\frac{\hbar^2}{8\pi^2 m} \frac{\partial^2}{\partial x_i^2} + \sum_j -\frac{\hbar^2}{8\pi^2 M_j} \frac{\partial^2}{\partial X_j^2} + V(x_i, X_j) \right] u(x_i, X_j) v(X_j) \\ &= E u(x_i, X_j) v(X_j) + \sum_j -\frac{\hbar^2}{8\pi^2 M_j} \left(2 \frac{\partial u}{\partial X_j} \frac{\partial v}{\partial X_j} + \frac{\partial^2 u}{\partial X_j^2} v \right). \end{aligned} \quad (5)$$

If it were not for the last summation, this would be exactly the equation we wish ψ to satisfy. But it is not difficult to show that these terms are small. Thus for example with the last one, u depends on the X_j 's in very much the same way in which it depends on x_i , since it depends largely on the differences $(x_i - X_j)$, representing the coordinates of electrons with respect to

the various nuclei, which are the essential things in the electronic motion. Hence $\partial^2 u / \partial X_j^2$ is of the same order of magnitude as $\partial^2 u / \partial x_i^2$. But this quantity, multiplied by $h^2 / 8\pi^2 m$, is of the order of magnitude of the energy of one of the electrons, an appreciable fraction of the energy of the system. The term appears here, however, multiplied by $h^2 / 8\pi^2 M_j$, smaller in the ratio of m / M_j , and since M_j is thousands of times m , this means that these terms are much smaller than the others, and can be neglected. Thus approximately uv forms a solution of the problem, and we are justified in using the electronic energy as a potential function for the nuclei. But to a higher approximation, we cannot neglect these small terms. We can find their matrix components between the state we are interested in and all other states, differing in electronic and nuclear motions, and these components, though small, will be different from zero.

It is these components, of the term $\sum_j -\frac{h^2}{8\pi^2 M_j} \left(2 \frac{\partial u}{\partial X_j} \frac{\partial v}{\partial X_j} + \frac{\partial^2 u}{\partial X_j^2} v \right)$, which determine the rate of transitions between different electronic levels, which we have considered in this chapter, and which we have seen are important in problems of collisions, nuclear vibrations in metals, etc.

Problems

1. If a reaction rate doubles for 10° rise of temperature, at $T = 300^\circ$, find the activation energy, in volt-electrons per atom, and in kg.-cal. per gram-molecule.
2. If three atoms interact by valence forces, it can be proved that the following formula gives approximately the energy of the lowest state: $\sqrt{\frac{1}{2}[(\alpha - \beta)^2 + (\beta - \gamma)^2 + (\gamma - \alpha)^2]}$, where α , β , γ , are the energies of binding of the pairs 1 and 2, 2 and 3, 1 and 3, respectively, if in each case the third atom is removed to infinity, so that α , etc., are given as functions of the three interatomic distances r_{12} , r_{23} , r_{31} by curves of the nature of Fig. 77. This formula is used in constructing Fig. 78. Show that the formula approaches the correct limit as any one of the three atoms recedes to infinity. Show that a single atom approaching a molecule is repelled, by assuming atoms 1 and 2 to be at the equilibrium distance, forming a molecule, so that α is large and negative, β and γ much smaller and also negative, increasingly so as the third atom approaches, and expanding the square root in binomial series in the small quantities β and γ .
3. Find the energy for three hydrogen atoms on a line at arbitrary distances apart, using the formula of Prob. 2, and the hydrogen interaction energy from Prob. 4, Chap. XXXV.

4. Taking the energy expression of Prob. 3, let the distances r_{12} and r_{23} be equal, so that the corresponding point is on the 45-deg. diagonal of Fig. 78, on which on account of symmetry the pass is located in this case. Compute energy as a function of r_{12} or r_{23} , and find graphically the energy of the pass (the minimum of this curve). Compare with the energy at the bottom of either valley, and so find the activation energy of the reaction in which a hydrogen atom approaches a hydrogen molecule, knocks off one of the atoms from the molecule, and itself becomes bound.

5. Suppose that two normal atoms collide, and we are interested in knowing whether either one can become excited. Let their kinetic energy be the average value corresponding to T . What would the temperature T have to be if the energy of excitation was 0.1 electron volt? 1 volt? 10 volts?

6. Show that two normal atoms colliding with sufficient kinetic energy may be bound to form an excited molecule under certain conditions, by the inverse process to predissociation, and describe by a diagram the necessary relations of electronic energy levels and excitation energy.

7. The collision of electrons with ions of a metallic lattice is as if the electrons were stopped after traveling a certain time T . Treating the electrons by classical mechanics, assume there are N per unit volume, all accelerated by the external field, and each stopped after traveling the time T , being reduced to rest, but immediately starting to speed up again. Find the mean velocity of each electron, and the electric current carried by all the electrons. Show that this is proportional to the external field, so that Ohm's law holds, and compute the specific resistance in terms of N and T , showing that the specific conductivity is $N \frac{T}{2} \frac{e^2}{m}$, where e and m are the charge and mass of the electron.

8. Assuming that copper has one electron per atom acting in metallic conduction, and taking the observed conductivity, find the mean time T between collisions, using the formula of Prob. 7.

CHAPTER XXXIX

ELECTRONIC INTERACTIONS

For some time we have been treating the motions of atoms and molecules, under the influence of a potential energy function which was really simply the energy of the electrons if the nuclei were considered as held fixed. We have derived many interesting results, but we have merely assumed the form of the potential energy function, or have discussed it by semi-empirical methods, rather than deriving it directly from wave mechanics. It is time now to return to this fundamental problem of electronic motion, with fixed nuclei, and to try to derive the many results we have used regarding the energy as a function of nuclear position. Even with atoms we have given so far only the roughest sort of approximation, and there are many features of both atomic and molecular structure, as for instance the structure of multiplets, and the proper formulation of the exclusion principle, which we have not touched. Our discussion will be carried out in general by means of the perturbation theory, and we shall apply it in a general way, so that it can be specialized for either atoms or molecules. A considerable part of our work must be devoted to the electron spin and its relation to the exclusion principle, for we have seen that this is fundamental in the theory of valence. The starting point will be the discussion of the many-body problem which we have already given in Chap. XXXIV. There we saw that we could get an approximate solution of the problem if we assumed that each electron, instead of being acted upon by all the others, was acted on instead by a force which depended on its position alone, this force being equal to the average force which the other electrons would exert, averaged over their motion. In this case, each electron is independent of the others, as far as its equations of motion are concerned, Schrödinger's equation can be separated, and the final wave function can be written as a product of functions of the various electrons: $U = u_1(x_1y_1z_1) \cdot \cdot \cdot u_n(x_ny_nz_n)$, where $u_1(x_1y_1z_1)$ is the one-electron function for the first electron, etc.

281. The Exclusion Principle.—The wave function we have just written down is a solution of the problem connected with the Schrödinger equation

$$\left[\sum_i -\frac{\hbar^2}{8\pi^2m} \nabla_i^2 + V_i(x_i y_i z_i) \right] U = EU, \quad (1)$$

which results from the separate equations

$$\left[-\frac{\hbar^2}{8\pi^2m} \nabla_i^2 + V_i(x_i y_i z_i) \right] u_i(x_i y_i z_i) = E_i u_i(x_i y_i z_i),$$

where

$$E_1 + \dots + E_n = E.$$

But now there is one special feature when all the particles $1 \dots n$ are electrons, and hence just alike. This feature is that each must move in just the same force field. In other words, interchanging the position of two electrons must leave the potential energy unchanged. The reason is plain: interchanging two electrons can make no physical difference, on account of their complete identity. In other words, we should write for the potential energy, not $V_i(x_i y_i z_i)$, but merely $V(x_i y_i z_i)$. And the various one-electron functions $u_i(x_i y_i z_i)$ are simply solutions of the same problem, but in general connected with different quantum numbers. Thus in a problem of atomic structure, each electron moves in the same central field, but different electrons have different quantum numbers.

An immediate result of the identity of electrons is that, as soon as we have found one solution of the Schrödinger equation above, we have likewise found many other solutions. For suppose that we take one solution, and then permute the quantum numbers in any arbitrary way among the electrons, we shall still have a solution. By this we mean that, if $u_a(x_1)u_b(x_2)$ is a solution of a two-electron problem, a and b referring to two quantum numbers, 1 and 2 to the coordinates of two electrons, then equally well $u_b(x_1)u_a(x_2)$ will be a solution; and so on for more complicated cases. Further, each of these wave functions will correspond to the same energy, so that we shall have a problem of degeneracy. And in such a case, we know that the correct solution is generally a combination of the degenerate solutions. From Chap. XXXII we know how to take care of such a situation. We shall have $n!$ different wave functions; for the n sets of quantum numbers can be permuted among

the electrons in $n!$ ways (unless some sets of quantum numbers are counted more than once, in which case there will be fewer wave functions). Thus we must solve a perturbation problem between these $n!$ functions, which we may number $1 \dots N$, where $N = n!$ and denote by $U_1 \dots U_N$. Then we find the matrix components of energy H between these functions, and in terms of these components, we set up the set of N simultaneous equations

$$\sum_m (H_{km} - E_p \delta_{km}) S_{mp} = 0, \quad k = 1 \dots N, \quad (2)$$

where $\delta_{km} = 1$ if $k = m$, 0 if $k \neq m$. The operator H is the whole energy, involving interactions between electrons, rather than the approximate one used in defining the u 's. These equations determine the coefficients S such that the linear combinations

$$\sum_m S_{mp} U_m \quad (3)$$

represent the correct wave functions after applying the perturbation, the one we have written being the p th perturbed wave function. In order to solve these equations, we have found that the determinant of N rows and columns formed from the quantities $(H_{km} - E_p \delta_{km})$, taking all N values of k and m , has to be zero. This gives an equation, called a secular equation, for E_p , having N roots giving the N energy levels, which we number by the index p from 1 to N . We should carry out this process in this case. Unfortunately it is too difficult to do, but fortunately a simplification is introduced by the exclusion principle which renders it easy to handle.

If we were able to solve the problem of degeneracy, we should find N linear combinations of the original products, $\sum_m S_{mp} U_m$, where $p = 1 \dots N$, each of which was a solution; that is, H operating on any one of these combinations would give just a constant times this combination itself. This is not true of the original products: $HU_m = H_{1m}U_1 + \dots + H_{Nm}U_N$, involving all the functions, not just the single one U_m , since H does not have a diagonal matrix with respect to the approximate functions U . But now if by some other method we can set up a combination of u 's which has the property that H operating on

it gives a multiple of itself, we shall immediately know that this combination is a solution of Schrödinger's equation. Fortunately we can do this in one case, and it turns out that this is the only case we are interested in. Suppose we set up the determinant

$$\begin{vmatrix} u_1(x_1) & u_1(x_2) & \dots & u_1(x_n) \\ u_2(x_1) & \dots & \dots & u_2(x_n) \\ \dots & \dots & \dots & \dots \\ u_n(x_1) & \dots & \dots & u_n(x_n) \end{vmatrix}. \quad (4)$$

This determinant by definition is a linear combination of all possible products of the form $u_1(x_1) \dots u_n(x_n)$, obtained by permuting the quantum numbers in all the possible N ways, each having a coefficient $+1$ or -1 according as an even or odd number of interchanges of rows or columns was necessary to bring the desired term to the principal diagonal of the determinant. In other words, it is a linear combination $\sum_m S_{mp} U_m$,

in which all coefficients S are ± 1 . And we can show that this particular combination actually has the property that H operating on it gives a multiple of itself (that is, that it does not have any nondiagonal matrix component to any other linear combination of the U 's). To do this, we must first note that the determinant is antisymmetric in the coordinates $x_1 \dots x_n$. By that we mean that if we interchange two coordinates with each other, as x_i and x_j , the whole function changes sign, but is otherwise unchanged, retaining the same magnitude. In other words, interchanging the position of two electrons makes only a change of sign in the wave function. The way in which we see that this particular function is antisymmetric is from a property of the determinant: to interchange two coordinates means to interchange the corresponding columns in the determinant, and there is a theorem of determinants stating that this interchange merely multiplies the determinant by -1 . The reason is simple: each product entering into the expansion of the determinant is unchanged numerically by the interchange of columns, but in each case one more or one less permutation is required to reach a given product than was required before, multiplying all factors by -1 , according to the rule of sign stated above. As a simple example, we start with the determinant

$$\begin{vmatrix} u_1(x_1) & u_1(x_2) \\ u_2(x_1) & u_2(x_2) \end{vmatrix} = u_1(x_1)u_2(x_2) - u_2(x_1)u_1(x_2),$$

and interchange columns, obtaining

$$\begin{vmatrix} u_1(x_2) & u_1(x_1) \\ u_2(x_2) & u_2(x_1) \end{vmatrix} = u_1(x_2)u_2(x_1) - u_1(x_1)u_2(x_2),$$

the same as the other but with opposite sign. Thus our linear combination is antisymmetric. It is easy to see that it is the only possible antisymmetric linear combination.

We now know, if we call our antisymmetric determinant D , that HD must be a linear combination of all functions U . But HD must be antisymmetric; for D is antisymmetric, and H , the energy, is symmetric, being entirely unchanged by interchange of two electrons, so that HD in turn will be changed only in sign. In other words, HD is an antisymmetric linear combination of the products U , and the only such combination, as we have just seen, is the determinant D itself, or at most a constant times the determinant. Hence we have shown that

$$HD = \text{constant} \times D,$$

or that D has the property we desired, of having no nondiagonal matrix components to other linear combinations of the U 's. We must not suppose that we have found an exact solution of Schrödinger's equation, though our description might indicate this; for H will have nondiagonal components between D and other antisymmetric functions formed from one-electron wave functions u of different quantum number from those used here. It is only within our group of N unperturbed wave functions that we have eliminated nondiagonal terms.

Out of all the N linear combinations of the N unperturbed wave functions, we have found just one which satisfies Schrödinger's equation. This seems like a rather small beginning toward the task of finding all N combinations, which we should obtain by solving the secular equation. But now the exclusion principle enters; and its statement is at first sight quite different from what we have become accustomed to. It is:

The only wave functions allowed in nature are those antisymmetric in all electrons.

This principle, as we have before pointed out, cannot be at present deduced from any results of wave mechanics, but must serve as a separate postulate. We can see at once, however,

that it is a consistent postulate, in the sense that if the universe were once set up with antisymmetric wave functions, it would always stay so. For Schrödinger's equation involving the time

is $H\psi = -\frac{\hbar}{2\pi i} \frac{\partial \psi}{\partial t}$, giving the time rate of change of ψ . If

now at a given instant ψ is antisymmetric, then $H\psi$ must also be antisymmetric, and hence the increment of ψ in time dt is also antisymmetric. Since the sum of two antisymmetric functions is itself antisymmetric, ψ at time $t + dt$, which is the original ψ plus its increment, will also be antisymmetric, or this property of antisymmetry does not change with time.

282. Results of Antisymmetry of Wave Functions.—The antisymmetry of the wave function, which we have just stated, results immediately in the ordinary form of the exclusion principle, the fact that no two electrons may have the same set of quantum numbers. For suppose that two of the one-electron wave functions, say u_i and u_j , were equal, as they would be if they corresponded to the same quantum numbers. Then we should have a determinant with two equal rows, and such a determinant is always zero, as we can see from the fact that interchanging these equal rows must surely leave the determinant unchanged, and yet interchanging two rows of any determinant must change its sign, inconsistent requirements unless the determinant is zero. Hence no antisymmetric wave function can be set up unless all electrons have different quantum numbers.

As a result of the exclusion principle, any particular set of electronic quantum numbers, and hence of wave functions $u_1 \dots u_n$, is connected with but one wave function for the whole system, instead of the $n! = N$ functions we at first had to consider. This greatly simplifies problems of electronic structure. There is one point connected with it, however, which is at first paradoxical. We can no longer speak of the quantum numbers of a particular electron. Each electron behaves just the same as any other electron. The quantum numbers refer merely to the one-electron wave functions from which we construct our antisymmetric wave function. We can visualize this situation if we think of an atom, with tightly bound K electrons, and a loosely bound valence electron. The same electron which acts at one time as valence electron may sometimes go near the nucleus and act like a K electron, but at the same time another electron will have changed place with it, and will now be acting

as valence electron. A similar process of interchange takes place in molecules. For example in H_2 , one cannot say that one definite electron is attached to one atom, the other to the other, for the electron which at one time is on one atom will at another time be on the other, with a corresponding change of the second electron. This process of electronic interchange is intimately connected with the formation of valence bonds, and is a very widespread phenomenon.

283. The Electron Spin.—In Chap. XXXV, it was stated that electrons have spins, as if they were permanent magnets, and that these magnetic moments are allowed to be oriented in only certain directions. For the present purposes, we can state the rule regarding their orientation in the following way: We may pick out some arbitrary direction in space, and then may postulate that each spin can be oriented either parallel or opposite to this direction, but not at an angle to it. The spin has angular momentum $\frac{1}{2} \frac{h}{2\pi}$, as if it had $l = \frac{1}{2}$, and correspondingly there are the two possible orientations $m = \pm \frac{1}{2}$, parallel or opposite to the axis. The spin may now be considered a little like a coordinate: four, not three, quantities are needed to describe the situation of an electron, its x coordinate, its y , its z , and its spin. The coordinates x, y, z are capable of taking on any value; but the quantity determining the spin, which we may take as the component of spin along the chosen direction, can have only two values, $+$ or $-$ the magnitude of spin itself. Our one-electron wave function should now depend not merely on x, y, z , but also on the spin. Since there are only two possible values which the spin can take, the wave function needs to be determined only for these two values, which we can symbolize by $+$ and $-$. We have, then, $u(x, y, z, \text{spin})$, defined only when spin is one of the two values symbolized by $+$ or $-$. In other words, we have $u(x, y, z, +)$ and $u(x, y, z, -)$.

The spin, as we have seen, behaves like a coordinate. But at the same time, it also acts like a quantum number, and this is apt to be rather confusing. Let us consider an electron in a central field. The three quantum numbers with which we are familiar are the total quantum number n , the azimuthal quantum number l , and the quantum number m . Of these, l measures the total angular momentum on account of the rotation of the electron in its orbit, in units of $h/2\pi$, and m measures the projection

of this angular momentum along a fixed axis, the z axis. But now the electron has an angular momentum on account of its spin, which proves to be $\frac{1}{2} \frac{h}{2\pi}$ in magnitude. This spin, as we have just seen, can be oriented in two ways with respect to a fixed axis, either along it or opposite to it. It thus appears that this spin angular momentum should likewise have quantum numbers similar to the orbital angular momentum, one representing its total magnitude (which, being always $\frac{1}{2}$, need not be specially considered, since the spin angular momentum, unlike the orbital angular momentum, never changes its magnitude), and the other its projection along the z axis (which can be either $+\frac{1}{2}$ or $-\frac{1}{2}$ units). Suppose this latter quantum number, determining the projection of spin along the axis, and capable of taking on just the two values $+\frac{1}{2}$ and $-\frac{1}{2}$, be called m_s . Then to specify the stationary state of an electron, we must give the four quantities n, l, m, m_s . As a matter of notation, it is often convenient to use the name m_l instead of m for the component of l along the axis, so that our four numbers are n, l, m_l, m_s . And the wave function should properly carry these four numbers as subscripts: $u_{n,l,m_l,m_s}(x, y, z, \text{spin})$.

We are now prepared to consider the physical meaning of the functions $u_{n,l,m_l,m_s}(x, y, z, +)$ and $u_{n,l,m_l,m_s}(x, y, z, -)$. The square of the first gives the probability that, if the quantum numbers are n, l, m_l, m_s , the coordinates will be $x, y, z, +$. Suppose that $m_s = \frac{1}{2}$. Then we know that the spin must be along the $+$ axis. In this case, there is no probability that the spin is along the $-$ axis, for we have information to the contrary. Thus $u_{n,l,m_l,m_s}^2(x, y, z, -)$ must be zero, since it measures the probability that the spin is along the $-$ axis. On the other hand, there is certainty that the spin is along the $+$ axis, so that $u_{n,l,m_l,m_s}^2(x, y, z, +)$ merely gives information about the distribution in x, y, z , or reduces to the ordinary function of x, y, z . A similar situation holds if $m_s = -\frac{1}{2}$ and the final result is

$$\begin{aligned} u_{n,l,m_l,m_s}(x, y, z, +) &= u_{n,l,m_l}(x, y, z) \text{ if } m_s = \frac{1}{2} \\ &= 0 \text{ if } m_s = -\frac{1}{2} \\ u_{n,l,m_l,m_s}(x, y, z, -) &= u_{n,l,m_l}(x, y, z) \text{ if } m_s = -\frac{1}{2} \\ &= 0 \text{ if } m_s = \frac{1}{2}, \end{aligned}$$

where $u_{n,l,m_l}(x, y, z)$ is the ordinary solution of Schrödinger's equation without spin. It is easy to see that these can be combined in the statement

$$u_{n,l,m_s}(x, y, z, \pm) = u_{n,l,m_l}(x, y, z) \delta(m_s, \pm \frac{1}{2}), \quad (5)$$

where $\delta(m_s, a) = 1$ if $m_s = a$, 0 if $m_s \neq a$. The wave function is then separated, a function of x, y, z times a function of spin, and the latter has this peculiar form δ . The separation is natural, since the energy does not depend on the spin. If we were including magnetic terms in our energy, which we have so far neglected, we should find that the magnetic moment of the spin actually does exert a small force, resulting in a small term in the energy, and when this is considered the separation is no longer possible.

284. Electron Spins and Multiplicity of Levels.—Suppose we have two electrons. The spin of each is $\frac{1}{2}$ unit of angular momentum, and can be oriented in either of two ways, parallel or opposite to the z axis. Thus we can have the following possibilities:

	1st	2d	Sum
I	$+\frac{1}{2}$	$+\frac{1}{2}$	+1
II	$+\frac{1}{2}$	$-\frac{1}{2}$	0
III	$-\frac{1}{2}$	$+\frac{1}{2}$	0
IV	$-\frac{1}{2}$	$-\frac{1}{2}$	-1

In other words, the total angular momentum of spin along the z axis, the sum of the two, has the possibility of being 1, 0, 0, -1. But there is another way of interpreting this. We may consider that the total angular momentum is the vector sum of the separate angular momenta of the two spins. If these are parallel, the sum is 1; if they are opposite, the sum is zero. With any intermediate angle, the result is between zero and unity. But such vector additions of angular momenta prove to occur often in quantum theory, and when they do, the vector sum is always quantized; that is to say, it has only the possibility of a discrete set of values, differing by unity from each other. Thus in this case the only possibilities for the total angular momentum are 0 and 1. The quantum number S is applied to this total angular momentum (s for spin, capital letter because it is a sum over several electrons, rather than for a single electron). Next, any angular momentum is allowed only certain quantized orientations in space, as the orbital angular momentum of a hydrogen electron is. In Fig. 68, we saw how an angular momentum l was allowed to have only the components m along the z axis, where $m = l$,

$l-1, l-2, \dots -l$. This law is also general, so that we see that our angular momentum 0 can have only the component 0 along the axis, whereas the other one 1 can have components 1, 0, -1 . For the state with $S = 0$, then, we have but one level, while for the state with $S = 1$ we have three levels.

Now the orientation of the vector S in space is a process involving but very small energy. The only forces on S prove to be magnetic forces, since the spin carries with it a magnetic moment, the motions of the electrons produce a magnetic field, and the relative orientation of the two affects the magnetic energy. This energy is small, however, so that the three levels of $S = 1$ lie close together, and form what is called a triplet. Similarly for $S = 0$ we have but one level, a singlet. On the other hand, we shall soon see that levels with different values of S generally lie far apart, with large energy separation. The effect of spins, then, is to produce multiplets, groups of levels close together, with considerable separation between multiplets. To verify these facts, let us compute the energy levels in the case of two electrons.

285. Multiplicity and the Exclusion Principle.—We have already considered one form of degeneracy inherent in our method of setting up an approximate solution of the wave equation: the exchange degeneracy, arising because it made no difference in the physical situation if two electrons exchanged positions. Now we must consider a second type: spin degeneracy, arising because (to the approximation to which we can neglect magnetic energy) it makes no difference in the energy which way a spin is oriented. In the last section, we have set up four combinations for the spins of two electrons. To each of these corresponds an antisymmetric wave function; for instance, to the second one,

$$\frac{1}{\sqrt{2}} \begin{vmatrix} u_{n_1, l_1, m_{l_1}, +\frac{1}{2}}(x_1) & u_{n_1, l_1, m_{l_1}, +\frac{1}{2}}(x_2) \\ u_{n_2, l_2, m_{l_2}, -\frac{1}{2}}(x_1) & u_{n_2, l_2, m_{l_2}, -\frac{1}{2}}(x_2) \end{vmatrix}, \quad (6)$$

where x_1 symbolizes the four quantities $x_1 y_1 z_1$ spin₁, and where the factor $1/\sqrt{2}$ simplifies the normalization. This determinant stands for the situation in which one electron is in the state with orbital quantum numbers $n_1 l_1 m_{l_1}$ and spin $+\frac{1}{2}$, the other in the state with quantum numbers $n_2 l_2 m_{l_2}$ and spin $-\frac{1}{2}$. There is one interesting fact which we may at once deduce from these determinants, and that is concerning the exclusion principle.

Suppose that our two electrons have the same orbital quantum numbers, so that $n_2 = n_1$, etc. Then the determinant in which both electrons have + spins, or in which both have - spins, is necessarily zero, for both electrons are then entirely alike, and the two rows of the determinant are alike and the determinant vanishes. But the determinant corresponding to spins + and - does not vanish; in this, the two electrons differ in spin, and so the exclusion principle does not forbid them to have the same orbital quantum numbers. To see this, we need only expand the determinant, which, writing the spin wave functions explicitly, is

$$u_{n,l,m_l}(x_1y_1z_1)u_{n,l,m_l}(x_2y_2z_2)[\delta(\frac{1}{2}, \text{spin}_1)\delta(-\frac{1}{2}, \text{spin}_2) - \delta(-\frac{1}{2}, \text{spin}_1)\delta(\frac{1}{2}, \text{spin}_2)] \quad (7)$$

The second factor is not zero; if the two spins are opposite, it is either +1 or -1, so that the wave function, as far as it depends on the coordinates, is simply the product of the functions of the two electronic coordinates. We readily see that the other determinant, corresponding to the first electron having a - spin, the second +, is just the same, except for a difference of sign, a trivial matter. Thus in this case of two electrons with the same quantum numbers, only one out of the four levels remains. This is clearly the singlet level. We then have the following very significant result:

Two electrons in general lead to a singlet and a triplet; but if they have the same orbital quantum numbers, they have only a singlet level.

The exclusion principle, in other words, can act to exclude certain multiplets, while permitting others. This proves to be a very important result in spectroscopy, since often a great many of the multiplets which would be allowed by the formal rules are excluded, simplifying the spectrum greatly. But the most important result is in the periodic table, and in other places where certain configurations are excluded entirely. Thus suppose we tried to have three electrons all with the same orbital quantum numbers. Then we simply could not choose their three spins so that all three would be different. The best we could do would be to have one +, two -, or *vice versa*. We should then inevitably have two electrons with just the same wave function, two equal rows in the determinant, and a wave function of zero. In other words, *no more than two electrons can have the same orbital wave*

function. And if two have the same wave function, they must have opposite spins, and hence form a singlet.

286. Spin Degeneracy for Two Electrons.—Let us avoid difficulty with the exclusion principle by assuming that our two electrons have different orbital wave functions, which to save writing we may symbolize by a and b . Similarly we shall symbolize the coordinates by 1 and 2, so that we can write a one-electron wave function of electron 1 in orbit a , with $+$ spin, as $a^+(1)$. Now we have our four combinations of spins, and each of these yields a different wave function. We have, then, a problem of degeneracy between these four functions, and we must set up the secular equation for this fourfold degeneracy, and solve it. The first step is to find the matrix components of the energy between the wave functions. And here a simplifying result appears, which we shall first prove: the matrix component of the energy (if we neglect magnetic terms) is zero unless both initial and final states have the same total spin. We can prove this most easily from the general method of finding a matrix component. Since the spin acts as a coordinate, we must sum over its two possible values, just as we integrate over each coordinate, in obtaining matrix components. Thus we have, for the matrix component between the first and second functions of our tabulation,

$$\int dv_1 \int dv_2 \Sigma(\text{spin}_1) \Sigma(\text{spin}_2) \{ [a^+(1)b^+(2) - b^+(1)a^+(2)] \\ H[a^+(1)b^-(2) - b^-(1)a^+(2)] \}. \quad (8)$$

Now $a^+(1) = a(1)\delta(+, \text{spin}_1)$, etc.; further, since H does not include the spin, it leaves the δ functions unchanged when it operates. Hence we may write our matrix component as

$$\begin{aligned} & \int dv_1 \int dv_2 \Sigma(\text{spin}_1) \Sigma(\text{spin}_2) \\ & \{ [\delta(+, \text{spin}_1)\delta(+, \text{spin}_1)\delta(+, \text{spin}_2)\delta(-, \text{spin}_2)] \\ & \qquad \qquad \qquad a(1)b(2)Ha(1)b(2) \\ & - [\delta(+, \text{spin}_1)\delta(-, \text{spin}_1)\delta(+, \text{spin}_2)\delta(+, \text{spin}_2)] \\ & \qquad \qquad \qquad a(1)b(2)Hb(1)a(2) \\ & - [\delta(+, \text{spin}_1)\delta(+, \text{spin}_1)\delta(+, \text{spin}_2)\delta(-, \text{spin}_2)] \\ & \qquad \qquad \qquad b(1)a(2)Ha(1)b(2) \\ & + [\delta(+, \text{spin}_1)\delta(-, \text{spin}_1)\delta(+, \text{spin}_2)\delta(+, \text{spin}_2)] \\ & \qquad \qquad \qquad b(1)a(2)Hb(1)a(2) \}. \end{aligned}$$

But now $\Sigma(\text{spin}_2)\delta(+, \text{spin}_2)\delta(-, \text{spin}_2)$ is zero; for it equals $\delta(+, +)\delta(-, +) + \delta(+, -)\delta(-, -)$, each term containing a

factor zero. Similarly each of the four terms is zero, and the matrix component vanishes. The same thing is readily seen to occur always if the total angular momentum is different in the two configurations.

The only components of H which are different from zero are then the diagonal ones, and the component between the second and third states. Let us compute these. We number our four levels from I to IV, as in the table in Sec. 284, so that for instance the function (6) is labeled II. We denote the matrix component of the energy between states I and II as $(I/H/II)$, with corresponding symbols for other components, and the matrix component of unity between the same states is $(I/1/II)$. Then we have

$$\begin{aligned}
 (I/H/I) &= \int dv_1 \int dv_2 \Sigma(\text{spin}_1) \Sigma(\text{spin}_2) \\
 &\quad \{ [\delta(+, \text{spin}_1) \delta(+, \text{spin}_1) \delta(+, \text{spin}_2) \delta(+, \text{spin}_2)] \\
 &\quad \frac{1}{2} \begin{vmatrix} a(1) & a(2) \\ b(1) & b(2) \end{vmatrix} H \begin{vmatrix} a(1) & a(2) \\ b(1) & b(2) \end{vmatrix} \} \\
 &= \frac{1}{2} \int dv_1 \int dv_2 \left[\begin{vmatrix} a(1) & a(2) \\ b(1) & b(2) \end{vmatrix} H \begin{vmatrix} a(1) & a(2) \\ b(1) & b(2) \end{vmatrix} \right] \\
 &= [(ab/H/ab) - (ab/H/ba)], \tag{9}
 \end{aligned}$$

where by definition

$$\int dv_1 \int dv_2 a(1)b(2)Ha(1)b(2) = (ab/H/ab), \text{ etc.} \tag{10}$$

Similarly we have

$$\begin{aligned}
 (II/H/II) &= (III/H/III) = (ab/H/ab) \\
 (II/H/III) &= (III/H/II) = -(ab/H/ba) \\
 (IV/H/IV) &= (I/H/I) \tag{11}
 \end{aligned}$$

We may now write down our secular equation; but we note first that our functions I . . . IV may not be normalized and orthogonal. Thus we have

$$\begin{aligned}
 (I/1/I) &= \frac{1}{2} \int dv_1 \int dv_2 \left[\begin{vmatrix} a(1) & a(2) \\ b(1) & b(2) \end{vmatrix} 1 \begin{vmatrix} a(1) & a(2) \\ b(1) & b(2) \end{vmatrix} \right] \\
 &= [(ab/1/ab) - (ab/1/ba)], \tag{12}
 \end{aligned}$$

where

$$\int dv_1 \int dv_2 a(1)b(2)a(1)b(2) = (ab/1/ab), \text{ etc.} \tag{13}$$

Similarly the other components are like those of H , but with 1 substituted in place of H . We see that, if a and b are separately normalized and are orthogonal, $(ab/1/ab) = 1$, $(ab/1/ba) = 0$, so that

$$\begin{aligned} (I/1/I) &= (II/1/II) = (III/1/III) \\ &= (IV/1/IV) = 1 \\ (II/1/III) &= (III/1/II) = 0, \end{aligned}$$

so that our functions are normalized and orthogonal; but we shall sometimes meet cases where this is not true.

Now we can write the secular equation for the energy. This is

$$\begin{vmatrix} I-J-E & 0 & 0 & 0 \\ 0 & I-E & -J & 0 \\ 0 & -J & I-E & 0 \\ 0 & 0 & 0 & I-J-E \end{vmatrix} = 0, \quad (14)$$

where $I = (ab/H/ab)$, $J = (ab/H/ba)$, E is the energy. This determinant can be at once factored:

$$(I - J - E)[(I - E)^2 - J^2](I - J - E) = 0, \quad (15)$$

giving a double root $E = I - J$, corresponding to the two states with components $+1$ and -1 of spin along the axis, and giving the two roots of the quadratic, $I - E = \pm J$, or $E = I \pm J$, for the two states corresponding to no spin along the axis. We have, then, three roots equal to each other, $E = I - J$; and one different root, $I + J$. Evidently the first, having three wave functions, with components of spin $+1$, 0 , -1 , forms the triplet, and the other is the singlet. To the order of accuracy to which we are working, neglecting magnetic terms in the energy, the three levels of the triplet fall exactly together, but they are separated by a considerable amount from the singlet, the separation being $2J$, where J is an integral depending on the electrostatic forces in H , and therefore of considerable size. This verifies our earlier statement that the energy depended in an important way on the total spin S , but only very slightly on its orientation.

287. Effect of Exclusion Principle and Spin.—The present chapter has been devoted to the mathematical formulation of the exclusion principle, and the effect on it of the spin, and to the method of finding energy levels subject to the complications

introduced by these features of electronic interactions. In the next chapter we shall make several physical applications of these ideas. For the present, we shall merely summarize what we have done, and briefly point out its importance. We first showed that the identity of electrons produced a degeneracy if we made approximate wave functions out of products of one-electron functions, on account of the possibility of exchanging electrons without making physical change in the system. We discovered, however, that out of the many perturbed wave functions allowed mathematically as linear combinations of these unperturbed ones, but one occurred in nature, the function which was antisymmetric in the coordinates of all electrons. This function had the property that it allowed no two electrons to have the same quantum numbers, the ordinary exclusion principle, but its importance extends much farther. Next we considered the spin, which had two possible orientations for each electron. This led to a new degeneracy, since each electron could have two possible spins, so that n electrons had 2^n possibilities (four possibilities for two electrons). We found that these 2^n levels broke up into groups, or multiplets, characterized by the total spin angular momentum, and such that all levels of a multiplet had the same energy, if we neglected magnetic effects, while different multiplets were separated widely from each other. This separation of multiplets is a result of the antisymmetry of the wave function, as we see if we look back over the argument, a result quite apart from the actual exclusion of certain levels, but equally or perhaps even more important. As we shall see when we analyze it mathematically, we have spoken of it in an earlier chapter as the effect of the exclusion principle on valence. The term $\pm J$, depending on this effect, will prove to be the term in the molecular energy which gives the valence binding or the repulsion, depending on whether the spins of the shared electrons are antiparallel (singlet state, energy $I + J$, binding, since J proves to be negative), or parallel (triplet, $I - J$, repulsion). It is somewhat paradoxical that this large and important effect of the spin on the energy can occur, when the spin exerts but negligible magnetic forces. These effects are not magnetic at all, but purely electrical, and they result simply because, on account of exclusion, the spin can exert a large influence on the wave function, the grouping of the electric charges, and the electrical energy. We can see

this most clearly from one property of the antisymmetric wave function which so far we have not pointed out: if we set the coordinates of two electrons equal in the determinant, the wave function vanishes, since interchange of the coordinates must change the sign, and yet can make no change when they are equal. This includes not merely space coordinates, but also spin. The result means the following physically: the probability is zero that two electrons of the same spin will be found at the same point of space (and small that they will be found near each other). On the other hand, the antisymmetry makes no objection to two electrons of opposite spin being close to each other. In other words, on account of this part of the exclusion principle, an electron of a given spin drives other electrons of the same spin away. And while this is not directly attributable to any force at all, still it can have a powerful effect on the electronic motions, and on the energy.

Problems

1. If the one-electron wave functions $u_1 \dots u_n$ are orthogonal and normalized, prove that

$$\frac{1}{\sqrt{n!}} \begin{vmatrix} u_1(x_1) & \dots & u_1(x_n) \\ \dots & \dots & \dots \\ u_n(x_1) & \dots & u_n(x_n) \end{vmatrix}$$

is normalized.

2. Show that a system containing an odd number of electrons always has even multiplets, as doublets, quartets, etc., while one with an even number of electrons has odd multiplets, singlets, triplets, etc.

3. Show that three electrons lead to a quartet and two doublets, on account of spin degeneracy. In case two of the electrons have the same orbital wave functions (that is, are equivalent), show that the quartet and one of the doublets are excluded.

4. In the problem of spin degeneracy of two electrons with orthogonal one-electron functions, as we have worked out, find the four final wave functions resulting from perturbation theory. Express these in terms of the δ functions for spin, and show that each one factors into a product of a function of coordinates and a function of spin. Show that the function of coordinates is proportional to $a(1)b(2) + a(2)b(1)$ for the singlet, $a(1)b(2) - a(2)b(1)$ for each of the levels of the triplet.

5. Discuss the spin degeneracy of two electrons, in the case where the one-electron functions a and b are not orthogonal, showing that the energies are

$$\frac{(ab/H/ab) \pm (ab/H/ba)}{1 \pm (ab/1/ba)},$$

the $+$ signs being for the singlet, the $-$ for the triplet.

6. Set up the perturbation problem of spin degeneracy for three non-equivalent electrons, and find the energy of the quartet terms, in the general case where the one-electron functions are not orthogonal.

7. If all wave functions are normalized and orthogonal, prove that the sum of all diagonal terms in the energy matrix equals the sum of all the perturbed energy values. To do this, expand the secular equation in the form $E^N - E^{N-1}(\text{coefficient}) + \dots = 0$, and also obtain a similar expansion for the factored solution of this equation, $(E - E_1)(E - E_2) \dots (E - E_N) = 0$, where $E_1 \dots E_N$ are the roots. Identify coefficients of the term in E^{N-1} in the two expressions.

8. Verify the result of Prob. 7 for the two-electron solution obtained in the text.

9. Using the method of Prob. 7, applied to the three-electron case with orthogonal one-electron functions, and the solution for the quartet energy found in Prob. 6, find the sum of the energies of the two doublet terms (or the mean energy of the two doublets).

CHAPTER XL

ELECTRONIC ENERGY OF ATOMS AND MOLECULES

In the last chapter we have seen how to set up the wave function of a system, subject to the exclusion principle and to spin, and how to find its energy levels by perturbation theory, taking account of the various degeneracies introduced. Space does not permit us to make complete applications of these methods to problems of atomic and molecular structure, but we shall indicate descriptively in the present chapter how the calculations are made, and the results they lead to.

288. Atomic Energy Levels.—In Chap. XXXIV, we have seen that the energy of an atom is primarily fixed by its configuration; that is, by the values of n and l for each electron. We have found approximate formulas for getting this energy, in any configuration, finding which electrons were tightly and which loosely bound, etc. And we have seen that in the low states of an atom, the electrons tend to be in the lowest levels possible, with two in the $1s$, two in $2s$, six in $2p$, etc., so that only the levels near the outside of the atom are unoccupied or only partly filled under ordinary conditions. Now we are prepared to go farther in understanding atomic structure.

An electron actually has not merely the two quantum numbers n and l , but also m_l and m_s , giving the orientation of orbital angular momentum and of spin, respectively. These only slightly affect the energy directly, only through small magnetic terms, which we have neglected. It was for this reason that in our earlier calculation of energy we could neglect them entirely. But these now introduce a degeneracy into the system, which we must consider in making a more accurate calculation of energy. We have already considered in the preceding chapter the spin degeneracy arising from m_s , connected with the different orientations of spin, but there is likewise an orbital degeneracy associated with m_l , arising from different orientations of orbital angular momentum. These two types of degeneracy become closely associated in atomic structure, and we must consider them

together. It is best to think of them first in terms of the vector model which we have used for discussing m_l for a single electron, and for m_s . Let us consider an atom with two electrons. Suppose, for illustration, that the electrons are nonequivalent p electrons (both having $l = 1$, but with different values of n). Then each has a total angular momentum of $\hbar/2\pi$ on account of its orbital motion, as given by l . These two vectors now form a vector sum, which we call L . If they are parallel, $L = 2$; if they are opposite, $L = 0$. As before, only values of L differing by integral values are allowed, or $L = 2, 1, 0$. Next, the two spin vectors also form a vector sum, S , which as we have seen can be in this case 1 or 0. Finally, L and S can be oriented in different ways with respect to each other, giving a resultant J , taking on all values with integral differences between $L + S$ and $|L - S|$. We have already seen that the energy depends in an important way on the value of S ; it also proves to depend in a similar way on L . The value of J , however, is unimportant, only the magnetic energy of coupling between spin and orbital motion depending on J . Hence the group of levels of the same L and S lie close together, and form a multiplet, but different multiplets lie comparatively far apart. It should be stated that this is a fairly special case, though an important one; cases may exist in which the magnetic energy is large, even sometimes larger than the energies dependent on L and S . It is only because here the coupling energies involved in forming L and S are large that it is correct first to set up these vectors, then to combine them to form J ; this case is then called $L - S$ coupling. But other types of coupling exist, and must often be considered.

In the matter of selection rules, and various other particulars, a multiplet of a many-electron atom with a given L behaves like a hydrogen level with the same value of l . For this reason, L is regarded as azimuthal quantum number, and there is a notation for levels like that in hydrogen, levels with L equal respectively to 0, 1, 2, . . . , being called S, P, D, F, \dots (large letters for complete atoms). A multiplet is indicated by a symbol as 3P , meaning a triplet P , with $L = 1$ (P), and $S = 1$ (triplet). And the separate levels of the multiplet, with $J = 2, 1, 0$, are denoted by $^3P_2, ^3P_1, ^3P_0$. To specify a level completely, one can give the configuration as well as the description of the term; as $1s2p^3P_2$, a level of a two-electron atom with one $1s$ and one $2p$ electron. Now we can see from our vector rules what multi-

plets arise from a given configuration. For instance, an s and a p electron give 1P and 3P ; two p 's give 1S , 1P , 1D , 3S , 3P , 3D , since L can be 2, 1, or 0, and S is 1 or 0. But here as in the case of the spin degeneracy alone, if the two electrons are equivalent, certain multiplets are excluded. With two equivalent p 's, for instance, the only multiplets which remain are the 1S , 1D , 3P . By classification of the levels, as we shall do in the next paragraph, this can be easily proved.

289. Spin and Orbital Degeneracy in Atomic Multiplets.—

As an illustration, we take the case of two p electrons, which to begin with we shall assume are not equivalent. The necessary information is given in the following table. In this we have indicated the orbital degeneracy completely, but not the spin degeneracy, since this follows the same arrangement as in the preceding chapter. In the first two columns we give the m_l 's of the two electrons, and in the third M_L , the sum of these two, giving the component of total orbital angular momentum along the z axis.

ORBITAL DEGENERACY FOR TWO p ELECTRONS

m_{l_1}	m_{l_2}	M_L	Non-equivalent	Equivalent
1	1	2	1 + 3	1
1	0	1	1 + 3	} 1 + 3
0	1	1	1 + 3	
0	0	0	1 + 3	1
1	-1	0	1 + 3	} 1 + 3
-1	1	0	1 + 3	
-1	0	-1	1 + 3	} 1 + 3
0	-1	-1	1 + 3	
-1	-1	-2	1 + 3	1

These are evidently the proper M_L 's to account for a D level ($M_L = 2, 1, 0, -1, -2$), a P ($M_L = 1, 0, -1$), and an S ($M_L = 0$), as demanded by the vector model. But now as a result of spin degeneracy, each one of these problems results in a singlet and a triplet, as indicated in the next column, so that we have just the set of multiplets already described as resulting from the configuration. In the case of equivalence of the electrons,

however, we have already seen that the triplet is not allowed if the two electrons have the same orbital wave functions. Thus, as indicated in the last column, for the two m_l 's equal respectively to 1, 1, or 0, 0, or -1, -1, only the singlet is allowed for equivalent electrons. Further, since the electrons are equivalent, the two arrangements 1, 0 and 0, 1 mean exactly the same thing, and yield only one singlet and triplet instead of two. Thus as we see from the last column we have a singlet level with $M_L = 2$, a singlet and triplet for 1, two singlets and triplet for 0, etc. And the only arrangement of multiplets which would yield this arrangement is 1S , 1D , 3P , as we have already stated.

We shall not carry out the computation of the energy of the multiplets, since this is rather a long and complicated task. It is not hard, however, to outline the process that must be used. In the first place, if the magnetic energy is neglected, it can be proved that the energy has no matrix components between states of different M_L . Let us assume that we have already solved, as in the last chapter, the separate problems of spin perturbation. Then in the case we are using for illustration, the degenerate perturbation problem of orbital degeneracy can be broken up into ten smaller problems: one for each of the five values of M_L , and each of the two allowable multiplicities. No one of these has more than a secular determinant of three rows and columns, and all are easily solved. The calculation of the integrals involved in the matrix components can be carried out. Since the one-electron wave functions are solutions of central field problems, they are functions of r times spherical harmonics of angle. The term in the energy which makes complication in the calculation of the matrix components is that in $1/r_{12}$, where r_{12} is the distance between the electrons. To handle this, we expand $1/r_{12}$ in spherical harmonics and inverse powers of r , and the integration over angles then resolves itself into an integration over products of spherical harmonics, which can be performed, leaving only an integral over functions of r . These integrals cannot be evaluated without knowing the functions of r contained in the wave function, and it is often convenient to leave them as undetermined parameters in the solution. Then we can get all matrix components in terms of a few of these parameters, and can solve the perturbation problem. When we do this, we find in the first place that all the levels of a multiplet come out automatically to have the same energy, as they should so long as we

neglect magnetic energies. Further, we find that the various multiplets are displaced from the center of gravity of all multiplets associated with the same configuration by amounts which are simple rational multiples of the various integrals, or parameters, which enter the problem. Thus, for instance, for two equivalent p electrons, there is but one parameter, and the 1S has energy 10 times the parameter, 1D 1 times it, and 3P -5 times it, all referred to the energy which we should obtain by the elementary theory neglecting degeneracy and multiplet structure. That is, 3P lies lowest, 1D next, 1S highest, and the energy separations are in the ratio of 2 to 3, a prediction which can be tested experimentally, even without knowing the numerical value of the parameter. In other cases, there are often several unknown parameters, so that we cannot predict immediately the relative values of the various separations, but still can get considerable information.

When now the magnetic energy of interaction between the magnetic moments of the spin and the orbital motion is included, this produces a further perturbation between the degenerate levels of each multiplet. For small magnetic energies this is not hard to work out. In this case, the energies of the various states of a multiplet follow simple rules, so that multiplets are groups of levels spaced in a regular fashion. When the magnetic energy is so large that multiplets spread enough to overlap other multiplets, however, the $L - S$ coupling no longer holds, and the situation becomes very complicated. It is no longer possible to classify into multiplets at all; an individual level may take on some of the properties of several different multiplets, its wave function being a linear combination of functions of these various multiplets. And even a greater complication often is present in actual cases: the multiplets connected with a given configuration may spread out so much that they overlap other configurations. Then even the distinction between configurations can become partly lost. We must solve a perturbation problem in which we take into account many configurations, not just one, and the final wave functions will be mixtures of these different configurations. The actual atomic structure, then, and the calculation necessary to describe it in detail, are very complicated.

290. Energy Levels of Diatomic Molecules.—Orbital degeneracy is of much less importance in molecules than in atoms, on account of the lack of spherical symmetry. The origin of

atomic orbital degeneracy is found in the fact that the energy of an electron is independent of the orientation of its orbit in space, so that a number of levels, having different orientations, correspond to the same energy. But with a molecule, different orientations, with different space arrangements of charge, will interact differently with the other atoms of a molecule, and hence will have different energies, and are to be counted as different configurations. The only special case is in a diatomic molecule, where we may take the axis of the molecule to be our preferred axis in space. Then each state is characterized by the component of angular momentum along this axis; and the other state with component just the negative of this will have exactly the same energy, corresponding in Bohr's theory to an orbit in which the electron was merely rotating in the opposite direction. Thus we shall have levels with a two-fold degeneracy, but no more. And in polyatomic molecules, even this degeneracy will generally be lost. It is worth noting also that molecules in which the atoms are in s states, and have no orbital angular momentum anyway, will necessarily have no orbital degeneracy. This includes many important actual cases.

Since it is unimportant, we shall neglect orbital degeneracy in molecular structure. The real complication comes in a different direction, as far as the theory is concerned. This is the question of the type of unperturbed one-electron functions to use. Two different methods have been used, each having some advantages, and we shall describe these methods, and point out their relations. The first is the method of Heitler and London, which starts out by assuming that the one-electron wave functions are just as in separated atoms, and applying perturbation theory to these functions. We shall begin by sketching the treatment of the lowest state of the H_2 molecule by this method, its best known application.

291. Heitler and London Method for H_2 .—Let us assume that we have two normal hydrogen atoms at distance R apart. We shall assume, with Heitler and London, that the one-electron wave functions are those of an electron moving about either hydrogen atom in the absence of the other, or simply the $1s$ functions of a hydrogen atom. Let the function representing an electron about the first nucleus be a , and around the second b . Now if we have two electrons, one in each of these two wave functions, we have a perturbation problem leading to spin

degeneracy, as we saw in the last chapter. There is no difficulty about equivalence of the electrons; for, while both wave functions are $1s$, they are about different nuclei, so that they correspond to different functions of coordinates. Hence we have a singlet and triplet level arising from the interaction. We can get their energies immediately, from the methods of the preceding chapter. It is easy to see that the one-electron functions are not orthogonal; a and b are both everywhere positive, so that it is impossible that $\int a(x)b(x)dv$ should be zero. Hence we must use the method for nonorthogonal functions developed in Prob. 5, Chap. XXXIX, and we find for the energies

$$\frac{(ab/H/ab) \pm (ab/H/ba)}{1 \pm (ab/1/ba)}, \quad (1)$$

where the $+$ signs are for the singlet, the $-$ for the triplet. These integrals are functions of the distance of separation R , and evaluation of them leads to the interatomic potential energy curves.

It is interesting actually to work out the values of the energy integrals. To do this, we note that H for the diatomic molecule (including all terms except nuclear kinetic energy, which in this method we leave out), is

$$H = \frac{-\hbar^2}{8\pi^2m}(\nabla_1^2 + \nabla_2^2) + \frac{e^2}{R} - \frac{e^2}{r_{a1}} - \frac{e^2}{r_{a2}} - \frac{e^2}{r_{b1}} - \frac{e^2}{r_{b2}} + \frac{e^2}{r_{12}}, \quad (2)$$

where first we have the kinetic energy of the two electrons, then the repulsion of the nuclei, the attraction of the electrons for the nuclei (r_{a1} representing for instance the distance between electron 1 and nucleus a), and finally the repulsion of the electrons for each other (r_{12} being the interelectronic distance). We also remember that, since a and b are solutions of atomic problems, we have

$$\left(-\frac{\hbar^2}{8\pi^2m}\nabla_1^2 - \frac{e^2}{r_{a1}}\right)a = E_0a,$$

with a similar equation for b , where $E_0 = -13.54$ volt-electrons is the binding energy of the hydrogen atom. Using these equations, we may compute Hab , obtaining

$$Hab = 2E_0ab + \left(\frac{e^2}{R} - \frac{e^2}{r_{a2}} - \frac{e^2}{r_{b1}} + \frac{e^2}{r_{12}}\right)ab, \quad (3)$$

where by Hab we mean $Ha(1)b(2)$. Now to find $(ab/H/ab)$, we must multiply this expression through by $a(1)b(2)$, and integrate over the coordinates. We assume a and b are normalized, though not orthogonal. Thus $\int a(1)b(2)a(1)b(2)dv_1dv_2 = 1$. We then have, from the first term of Hab , simply $2E_0$, the energy of the unperturbed atoms. The term in e^2/R represents the repulsive energy between the two nuclei. The next term is

$$-\int a^2(1)b^2(2)\frac{e^2}{r_{a2}}dv_{12} = -\int b^2(2)\frac{e^2}{r_{a2}}dv_2,$$

integrating over the coordinates of the first electron. But this is just the potential at the nucleus a of a charge $-e$ distributed according to the density function b^2 , around the nucleus b , multiplied by the charge e of nucleus a . In other words, this term (and the next one, which is analogous) represents the attractions of the nuclei of each atom for the electron of the other. Finally the last term is

$\int a^2(1)b^2(2)\frac{e^2}{r_{12}}dv_{12}$, or the repulsive interaction between a charge e distributed over the first nucleus according to the density a^2 , and another charge on the second nucleus with density b^2 . The four terms taken together, then, represent the electrostatic interaction between the two atoms, each regarded as a nucleus and a charge distribution of negative charge surrounding it. They are, in other words, the penetration or Coulomb interaction which we have discussed in a previous chapter. There by qualitative arguments we showed that this interaction would give a net attraction at moderate distances, though at sufficiently small distances, on account of the nuclear repulsion e^2/R , the interaction always becomes repulsive. Our first result, then, is a formula for this penetration interaction.

The other term entering into the energy is the integral $(ab/H/ba)$. This is the same as $(ba/H/ab)$, and is obtained by multiplying $Ha(1)b(2)$ by $b(1)a(2)$, and integrating. In doing this, we encounter at once the integral $\int a(1)b(1)dv_1\int a(2)b(2)dv_2$, or the square of $\int a(1)b(1)dv_1$. Now the function a is large near the first nucleus, and falls down exponentially as we go away from it, becoming small at the second nucleus. Similarly b is large near the second, small at the first. The product then is small, since everywhere one or the other factor is small. The largest values may be assumed to come in between the two nuclei, where the two atoms overlap to an appreciable amount if

they are near enough together, so that both factors, representing the densities of each atom at such a point, are fairly large. The integral, then, will be a quantity small compared with unity, the principal contributions coming from the region of overlapping between the atoms. We can now examine the individual terms. The one in $2E_0$ will be simply $2E_0 \int a(1)b(1)a(2)b(2)dv_{12} = 2E_0(ab/1/ba)$. This, then, is just such that in the whole expression for energy we have the term

$$\frac{2E_0 \pm 2E_0(ab/1/ba)}{1 \pm (ab/1/ba)} = 2E_0,$$

the unperturbed energy. Next, the term in e^2/R is equal to $(\alpha e)^2/R$, if we write $\alpha = \int a(1)b(1)dv_1$, so that $\alpha^2 = (ab/1/ba)$, and as we have seen α is a number small compared with unity. This is the repulsion of two charges αe , one on each nucleus, for

each other. The next term is $-\int a(1)b(1)dv_1 \int a(2)b(2)\frac{e^2}{r_{a2}}dv_2 = -\alpha \int a(2)b(2)\frac{e^2}{r_{a2}}dv_2$, which is the potential energy of interaction

between a charge αe on nucleus a , and a charge of density $-a(2)b(2)$ distributed in the region of overlapping. This latter charge has the total amount $-\alpha e$. Similarly the next term is the attraction between a charge αe on the nucleus b , and $-\alpha e$ spread out in the region of overlapping, and the last term is the repulsion between two charges $-\alpha e$, each distributed over the region between the atoms. When now we compute these terms, we find that the attractions between the charges on the nuclei and the distributed charges between are more than enough to balance the repulsions between the charges on the nuclei, and between the distributed charges, and the net effect is a negative integral, at least at fairly large distances, giving attraction. If then in our energy expression we have the $+$ sign, the term $(ab/H/ba)$, being negative, will result in binding between the atoms. This is the valence attraction of which we have spoken before. We see that its physical interpretation is as follows: part of the electronic distribution from each atom moves into the region between the two atoms, forming an electron pair there. The amount of charge moving from each atom is $-\alpha e$, leaving a corresponding positive charge on each nucleus. The cause of the valence binding is the electrostatic energy of attraction between this concentration of negative charge in

the region between the atoms, and the residual positive charge left on each nucleus. If we use the $-$ sign instead, we obtain the repulsive energy level, discussed in Prob. 4, Chap. XXXV, the energy being repulsive because the term $-(ab/H/ba)$, which is a positive energy, is more than enough to counterbalance the Coulomb attraction. It is easy to see that this repulsive level is the triplet, corresponding to having the spins of the two electrons parallel, while the attractive level is the singlet, with the spins opposite. As we have pointed out before, this may be qualitatively connected with the exclusion principle, in that if the spins are opposite, the exclusion principle does not operate and the charges can overlap, while if they are parallel the charges cannot overlap, and a repulsion results.

292. The Method of Molecular Orbitals.—The other method which has been used for the discussion of molecular energy levels is one in which we take account of the fact that the actual electrons in a molecule are in a different field from that of a single atom, and try to find their one-electron wave functions subject to this field. Thus for H_2 , either electron may be near either nucleus. When it is near one, it is attracted by that nucleus, and when near the other by the other, so that it moves in a field with two attracting centers. The method we are now describing tries to find the wave functions of an electron in such a field, and uses these (sometimes called molecular orbitals) to build up a solution for the whole problem. It is not very easy to get exact solutions for the problem of two centers, but we can find an approximation fairly simply, by perturbations from the problem of one center. When the electron is near nucleus a , the actual field acting on it, in the molecule, is not very different from the field if the other atom were absent. Thus the wave function for a single electron, near nucleus a , must resemble our function $a(1)$ which we have used previously. On the other hand, when the same electron is near nucleus b , its wave function resembles $b(1)$. Thus the whole molecular orbital which we are seeking must have both properties at once of resembling a near the first nucleus, b near the second. We can try to build up a wave function as linear combination of these two, a and b , and in doing this we are led to a perturbation problem, each of these acting as an unperturbed wave function. Further, on account of the identity of the two nuclei, these two unperturbed functions correspond to the same energy, and the problem

is degenerate. When we solve the problem of degeneracy, we find easily that the two final wave functions are, except for a factor, $a + b$ and $a - b$. The first of these is symmetric in the two nuclei, while the second is antisymmetric. We show graphs of these two functions, taken by plotting values along a line joining the nuclei, in Fig. 80, where we see that the antisymmetric function has a node between the nuclei. Calculation of the one-electron energies of these two functions shows that the function without the node is more tightly bound. It is easy to see that the symmetrical function corresponds to an extra amount of charge between the nuclei, since the wave function there is twice as big as for either atom separately,

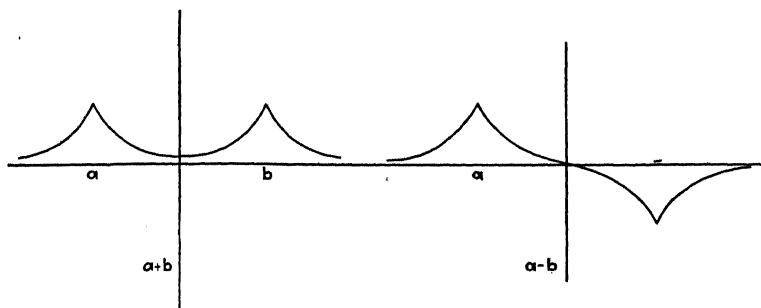


FIG. 80.—Symmetric and antisymmetric molecular orbitals. Figures represent values of the wave function at points on the line joining the centers of the atoms. Curve $a + b$ is symmetric, $a - b$ antisymmetric, where a and b are one-electron wave functions about the two nuclei.

and therefore its square, the density, is four times as great, or twice as much as for the two atoms separately. Thus there is a sort of interference effect between the waves, as in optics, where the amplitudes add, but the intensity does not. On the other hand, with the antisymmetric function, with its node in the center, there is less charge between the nuclei than for the two atoms separately.

Now that we have found the molecular orbitals, we remember that we have two electrons, each of which must go into one of the orbitals. In the lowest state, both electrons will be in the lowest orbital, with opposite spins (so as not to contradict the exclusion principle). Then there will be the extra concentration of charge between the nuclei which we have already pointed out in the symmetrical orbital, and which produces a binding between the atoms, as in the method of Heitler and London.

On the other hand, if both nuclei have the same spin, so that we have a triplet level, they cannot both be in the same orbital, but one must be in the higher, antisymmetric level, and this counteracts the effect of the attraction, and results in repulsion. It is a rather complicated thing actually to compute the energy by the method of molecular orbitals, and its use is more in qualitative discussion of the types of molecular structure, rather than for numerical computation. It can be shown, however, that it must lead to essentially the same results as the method of Heitler and London.

As an illustration of the sort of case where molecular orbitals are particularly useful, we may discuss the structure of the two molecules CO and N₂. Each of the atoms, in either of these molecules, contains two *K* electrons. In addition, either of the molecules has ten more electrons. Of these ten, two presumably act as 2s electrons about each of the nuclei, but the six others behave like molecular orbitals about the two nuclei. In these particular molecules, the nuclei are fairly close together, and 2*p* electrons are sufficiently extended in space so that they practically surround both nuclei. In this case, the 2*p* molecular orbitals in the field of two nuclei are not very different from 2*p* atomic orbits, there seem still to be six of them, and in the completed molecule there is one electron in each, resulting in a molecule which has a complete shell of 2*p*'s surrounding the whole thing, and therefore rather like a neon atom, small, inert chemically. This as a matter of fact is characteristic of the two molecules in question. In a case of this sort, evidently it would not be accurate to find the molecular orbitals by perturbation methods from atomic ones, as we did for H₂, but they should be found specially for the problem under discussion.

There is one problem for which the method of molecular orbitals is decidedly more convenient than that of Heitler and London, and that is the structure of metals. There a molecular orbital represents an electron which moves in the field of all other electrons and nuclei. Now this field, while it has great variations from point to point as we go from a position near a nucleus to one farther from nuclei, still is in general constant throughout the metal, showing only local fluctuations, unless an electric current is flowing, in which case, by Ohm's law, the field has a potential which varies slowly as we go through the metal. A molecular orbital in such a field, as we shall see in

the next chapter, while it shows the behavior of an atomic wave function near an individual atom, still varies through large distances as the wave function for a free electron would, moving in the averaged-out field. Now the interesting point is that we can easily set up such orbitals corresponding to electrons carrying a current, and it is by this method that electrical conductivity is described. The corresponding process is very difficult to treat by the method of Heitler and London.

Problems

1. Prove by the vector diagram that an S level, no matter what the multiplicity, has only one sublevel (one J value), and that a P level never has more than three sublevels.
2. Discuss by the vector diagram the levels arising from a p and a d electron; two nonequivalent d 's.
3. Prove that a closed shell of p electrons has no orbital or spin angular momentum, so that its state is 1S . Show that the same thing is true of any closed shell.
4. Prove that any configuration of electrons outside closed shells leads to the same set of multiplets that it would if the closed shells were absent. Thus prove that all alkali spectra are similar except for magnitude of the terms.
5. Work out the problem of orbital degeneracy of two equivalent d electrons, showing that the only allowed levels are $^1S^3P^1D^3F^1G$.
6. Prove that the vector diagram and the method of orbital degeneracy lead to the same set of levels for three nonequivalent p electrons. (Hint: in the vector method, first couple two of the l 's together to form a vector, and then couple the remaining l to this to form L . Proceed similarly with the spins.)
7. Prove that three equivalent p electrons lead to $^4S^2D^2P$.
8. Using the Heitler and London method, find an expression for the density of charge in the normal state of H_2 , as a function of position. Show that the density is greater in the region between the atoms than if we simply added the densities of the two atoms.
9. Using the molecular orbital $(a + b)/2$ (neglecting the fact that this is not exactly normalized), for H_2 , and an internuclear distance of 0.8 \AA , find the charge density at points in a plane containing the nuclei. Draw a diagram with lines of constant charge density, which would be circles surrounding the nucleus for a single atom, but show that in this case some of the lines surround both nuclei.
10. Draw a diagram similar to that of Prob. 9 for the charge density of the repulsive orbital $(a - b)/2$ for H_2 .

CHAPTER XLI

FERMI STATISTICS AND METALLIC STRUCTURE

For several chapters we have been dealing with the electronic structure of atoms and molecules, treating them by the perturbation theory applied to Schrödinger's equation, with the addition of the exclusion principle (antisymmetry of wave functions), and the electron spin. There is an alternative method, based directly on statistics, which in its present form is not capable of giving exact results, but which is very useful for qualitative discussions, and is not greatly in error when used numerically. This is the method of Fermi statistics. It is a method in which the exclusion principle is properly taken care of, but which treats the electronic motions, and Schrödinger's equation, only approximately. We may begin its discussion by treating the free electron theory of metals, one of the simplest applications.

293. The Exclusion Principle for Free Electrons.—Let us consider free electrons in a box, subject to the boundary condition that the wave function goes to zero on the boundaries. The problem is that of a perfect gas, if we neglect forces on the electrons, and we have already treated it in Chap. XXXV. We found there that the wave function for a single electron was $\sin \frac{n_1\pi x}{A} \sin \frac{n_2\pi y}{B} \sin \frac{n_3\pi z}{C}$, where n_1, n_2, n_3 were integers, A, B, C the sides of the box. The corresponding energy is

$$E = \frac{h^2}{8\pi^2m} \left(\frac{n_1^2\pi^2}{A^2} + \frac{n_2^2\pi^2}{B^2} + \frac{n_3^2\pi^2}{C^2} \right). \quad (1)$$

Let us consider the same electron in the phase space, and investigate its quantum conditions. The variables can be separated, and suppose we take the x coordinate and its momentum, drawing a phase space for these two variables (as in Fig. 81). In such a phase space, the path of a particle is represented by a line as $abcd$, where along ab the momentum is positive and constant, corresponding to motion along the box from left to right; bc , in which the momentum changes suddenly without

change of coordinate, represents the collision with the wall; cd represents the motion back in the opposite direction, and therefore with negative momentum, to the other wall, and da represents the collision with this wall. The ordinary quantum condition would state that $\oint p \, dq = nh$, or area $abcd = \text{integer times } h$. If p_x is the positive momentum along the path ab

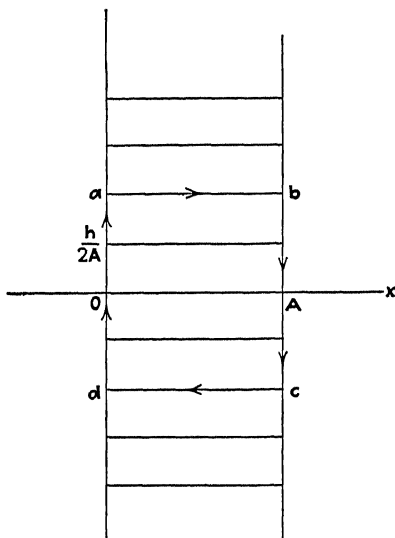


FIG. 81.—Phase space for free electron in box. The limits of the box are 0 and A , along the x axis. The rectangle $abcd$ represents the path of a particle, for which $\oint p \, dx = 2h$.

this evidently gives $2Ap_x = \text{integer times } h$, $p_x = \text{integer times } \frac{h}{2A}$. For the particular path indicated, the integer is evidently 2, the particle being in the second quantum state. It is easy to see that this relation between momentum and quantum number holds in the same way, with integral quantum numbers, for the wave mechanical solution. There the function $\sin \frac{n_1 \pi x}{A}$ can

be written as $\frac{1}{2i}(e^{in_1 \pi x/A} - e^{-in_1 \pi x/A})$, and when we multiply by the time factor, this gives the sum of two progressive waves traveling in opposite directions, representing the stream of particles going across the box and returning. For each of these streams, we can find the momentum; thus for the first term,

$\frac{\hbar}{2\pi i} \frac{\partial}{\partial x} (e^{in_1\pi x/A}) = \frac{\hbar}{2A} n_1 e^{in_1\pi x/A}$, showing that the x momentum is $n_1 \frac{\hbar}{2A}$, as before. Similarly the energy is at once seen, in terms of these values for p_x, p_y, p_z , to be $\frac{p_x^2 + p_y^2 + p_z^2}{2m}$, the classical value.

Now let us see how the exclusion principle operates, in the phase space. We may represent each electron by a separate point in the same six-dimensional phase space, and each one will move in such a path as $abcd$, at least when projected into the two-dimensional space associated with one coordinate. Now the exclusion principle states that only two electrons can have the same set of orbital quantum numbers, and these two must have opposite spin. In other words, only two electrons can have the same values of n_1, n_2, n_3 . This can be formulated in a statement that there is a certain maximum density of electrons in the phase space, which may not be exceeded. For associated with the quantum number n_1 , there is a definite area of the $x - p_x$ plane: the area between the path $abcd$ and that corresponding to the next smaller path, which from the quantum condition is h . Similarly associated with each of the two other quantum numbers is a two-dimensional area h , and when we put these together, we find a six-dimensional volume h^3 associated with a given set of n 's. In other words, the exclusion principle states that the maximum density of points corresponding to free electrons in the six-dimensional phase space is two electrons per volume h^3 , or $2/h^3$ electrons per unit volume, or a charge $-2e/h^3$ per unit volume. Since the six-dimensional density in phase space is the product of the ordinary density in space, by the density in momentum space, we see that this product cannot exceed a definite value. The denser the electrons are in ordinary space, the smaller must be the density in momentum space, therefore the greater must be the volume of momentum space occupied, and consequently the larger the maximum momentum, and kinetic energy, of the particles. Crowding electrons together in three-dimensional space, therefore, necessarily increases their kinetic energy, therefore requires work, and this effect, depending on the exclusion principle, is the one we have spoken of in connection with the repulsion of atoms and molecules.

294. Maximum Kinetic Energy and Density of Electrons.—

Two conclusions may be easily derived from the formulation of the exclusion principle in the phase space. First, let the density of electrons in ordinary space be determined. Then, even though the electrons are all in their lowest energy levels, the density of electrons in the momentum space cannot exceed the maximum allowed by the exclusion principle. As a result, the electrons must have kinetic energy, even at absolute zero of temperature, and we can easily compute the maximum kinetic energy which any of them must have. To do this, let us consider a three-dimensional momentum space. In this space, where p_x , p_y , p_z are the three variables, a surface of constant kinetic energy is a sphere: $E_k = \frac{p_x^2 + p_y^2 + p_z^2}{2m}$, a sphere of radius

$p = \sqrt{2mE_k}$. Now we assume that electrons are in the lowest possible energy levels, so that they have the smallest possible kinetic energy. In other words, the part of the momentum space within a sphere corresponding to the maximum kinetic energy, E_{\max} , will be filled up to its maximum allowable density with points, and no points will be found outside. We may then make the following equation: the actual density of particles in the ordinary space equals the integral of the density in phase space, integrated over the momenta. In other words,

$$\rho = \iiint (\text{density in phase space}) dp_x dp_y dp_z.$$

But the maximum density of charge in the phase space, as we have seen, is $-2e/h^3$. Further, in place of the integration $\iiint dp_x dp_y dp_z$, we may simply multiply by the volume of momentum space which is occupied, since the integrand is constant. This is the volume of the sphere of radius $\sqrt{2mE_{\max}}$, or $\frac{4}{3}\pi(2mE_{\max})^{3/2}$. In other words, we have

$$\rho = -\left(\frac{2e}{h^3}\right)\left(\frac{4}{3}\pi\right)(2mE_{\max})^{3/2}. \quad (2)$$

Solving for E_{\max} , we have

$$E_{\max} = \frac{1}{2m}\left(-\frac{3h^3\rho}{8\pi e}\right)^{2/3}. \quad (3)$$

This gives the maximum kinetic energy as a function of density of electricity, and shows as we expected that this kinetic energy

increases with density. In other words, as the electrons are forced closer together, the kinetic energy increases.

The second way of stating our result is the inverse: if the maximum allowable kinetic energy of electrons is somehow determined, then the density cannot be greater than the maximum value given by our equation. We shall shortly investigate a simple model of atomic structure based on this use of the theorem. There we assume electrons to be bound in the field of the nucleus, and apply this result to them. Surely no electron can be bound if its kinetic energy is great enough to allow it to escape from the center of attraction. This condition places a maximum on the possible kinetic energy of an electron in the atom. In turn, this determines a maximum density of electricity at every point of the atom, a density which is approximately reached in actual atoms.

295. The Fermi-Thomas Atomic Model.—We have just described a relation between maximum kinetic energy and maximum density of electrons. This has been proved only for free electrons; but since the law that the maximum density of electrons in phase space is $-2e/h^3$ holds even in a force field, we may expect our relation to be general (though not exact, since it does not take account of the fact that the quantum conditions are not the exact formulation of wave mechanics). Fermi and Thomas have applied this result to the problem of atomic structure. Suppose that we have a nucleus, surrounded by a cloud of electrons. Let the electrostatic potential at any point within this cloud be $V(r)$, so that the potential energy of an electron at the corresponding point is $-eV(r)$. The potential will go off to zero at large distances, since the atom is uncharged as a whole. Thus, if the total energy of an electron is negative, it cannot escape from the field, but if it is positive it can escape. We may be sure, then, that all electrons bound in the atom have negative energies, so that the maximum kinetic energy allowable for an electron at any point is $eV(r)$. This means that the density of charge at distance r from the center cannot exceed $-\left(\frac{8\pi e}{3h^3}\right)(2meV)^{3/2}$. Let us assume that the density has just this value. That gives us, then, a relation between charge density ρ and potential V . But of course there is another relation: the potential V is assumed to be produced by the charge itself, according to electrostatics, and in that case we have Poisson's

equation $\nabla^2 V = -4\pi\rho$. Since V is spherically symmetrical, the Laplacian can be written $\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial V}{\partial r} \right)$. Equating the two expressions for density, we finally have

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{dV}{dr} \right) = \frac{32\pi^2 e}{3h^3} (2meV)^{3/2}. \quad (4)$$

This equation must, of course, be solved subject to the condition that V for very small r behaves like the potential of the nucleus alone (thus bringing in the atomic number) and for large r approaches zero (thus determining that the electrons should be sufficient in number to balance the nuclear charge). It is a nonlinear differential equation for V , and cannot be solved except by numerical methods. It is easy to show, however, that by making a change of scale we can reduce the problem of arbitrary nuclear charge to a single equation, so that the problem can be solved once for all for all atoms. This has been done, and it is found that the resulting potential and charge density, while by no means accurately equal to the ones found by more elaborate methods, still are approximately correct, and good enough for a number of kinds of calculation. Unfortunately, the method is least accurate for the outer parts of the atom, where the density is small and actually the electrons do not have the maximum density allowed by the exclusion principle, and since these outer parts are most important in many applications, this method of Fermi and Thomas has not had as wide use as it otherwise might.

296. Electrons in Metals.—We have pointed out in another chapter that the electrons in a metal, though they are in a field which has intense local irregularities at the various atoms, still on the average are in an approximately field free space. The attractions of the nuclei are, of course, balanced by the repulsions of other electrons, and the electrons are largely free. We may thus approximately apply the ideas of maximum density of electrons, etc., developed in the present chapter. Let us see what picture of the electrons in a metal results from this.

There are an infinite number of possible stationary states for a free electron in a metal, corresponding to different values of momentum and kinetic energy. The energy levels are spaced in such a way that the energy is a quadratic function of the

quantum numbers. At absolute zero, the electrons naturally settle into the lowest energy levels they can, but just as in a single atom, the exclusion principle operates to prevent more than two electrons (with opposite spins) going into one stationary state. Thus some of the electrons will have to be in very high quantum states, and will have much kinetic energy. We have found the value of the maximum kinetic energy, and when we put in numerical values, assuming the number of electrons to be of the order of magnitude of the number of atoms in the metal, we find that this kinetic energy is of the order of magnitude of 10 volt-electrons. The electrons in a metal, then, are continuously moving about with these large energies, having speeds far in excess of those which would result from thermal agitation even at enormous temperatures, on ordinary statistics. They do not ordinarily carry any current, however, since as many go in one direction as in the opposite direction, so that the current cancels. In the presence of an external electric field, however, the electrons are accelerated, as shown in Chap. XXXVIII. As we saw there, this results in an electric current. The electrons gain only a small amount of energy and momentum in this acceleration, since they collide with the nuclei and lose their excess energy very soon, and the small drift velocity they acquire in the process corresponds to the net current we observe, which by Ohm's law is proportional to the field producing it.

It is interesting next to consider what happens to the metal as the temperature is raised from the absolute zero. The first effect is on the nuclear vibrations, and this affects directly the electrical conductivity. We have really not been entirely accurate so far in our description of the process of collision between the electrons and the nuclei. Since the electrons can be replaced by waves, the process is like the scattering of waves by a set of particles. Such a problem is met in optics, and there it can be shown, as in Sec. 184, that a medium with uniformly spaced particles does not scatter at all; the particles, like the atoms of a homogeneous transparent solid, affect the refractive index, but do not scatter the light in random directions. It is only when the particles show fluctuations of density, like the molecules of a gas, that they scatter, and scattering is proportional to the mean square deviations of the particles from positions of uniform spacing. In a similar way, electron waves are not scattered by uniformly spaced atoms, but only are

deflected if the atoms show nonuniformity in their arrangement. Now at the absolute zero, the atoms of a metallic crystal are uniformly arranged, so that the electrons are not scattered, and the resistance goes to zero, as it experimentally does. The zero point lattice vibration produces a nonuniform arrangement, it is true, but it can be proved that this does not add to the resistance. As the temperature increases, however, temperature agitation results in deviations from uniform arrangement of the atoms. The mean square deviations, and hence the scattering and the resistance, are proportional to the temperature, explaining in a very simple way the well-known experimental law giving the temperature coefficient of resistance.

In addition to this effect of temperature on nuclear vibrations, however, we may ask if there is any effect on electronic motions. There is such an effect, a small one. The electrons try to take up thermal energy of agitation. Those lying in low levels of the Fermi distribution, however, with kinetic energy much below the maximum kinetic energy of the electrons, cannot take up temperature energy. The reason is that in order to take any, their energies would have to increase enough so that they would have more kinetic energy than the maximum we have computed, since all energy levels below that are filled already, and this would require more energy than is available. The few highest electrons, however, have unfilled levels directly above them, and they can be raised to these levels by comparatively small additions of energy, which they can secure from temperature energy at ordinary temperatures. The situation is much as it is in atoms. The inner electrons of an atom cannot take part in temperature agitation, since the levels slightly above them are already occupied, and the least energy the electrons could take up would be enough to remove them entirely from the atom, an energy enormous relative to the amounts available at ordinary temperatures. On the other hand, the outer electrons can in some cases have temperature energy, for sometimes there are unoccupied levels lying only slightly above the highest occupied ones, which can be reached by addition of small amounts of energy.

It is interesting to draw diagrams of the change of electron distribution with temperature. At the absolute zero, we have seen that the momentum space is filled with electrons at a uniform density up to a maximum energy E_{\max} . The number

of electrons with energy between E and $E + dE$ is then proportional to the volume of momentum space between these energies, or proportional to $p^2 dp$, the shell between p and $p + dp$. Substituting E proportional to p^2 , this gives a shell of volume $\sqrt{E} dE$. In other words, the number of electrons per unit energy range is proportional to \sqrt{E} , up to E_{\max} , and above that it is zero. This is shown in Fig. 82, curve $oabc$, rising according to a square root law to E_{\max} , then dropping suddenly. At a higher temperature, there are fewer electrons with energy almost up to E_{\max} , but these now have energy slightly above

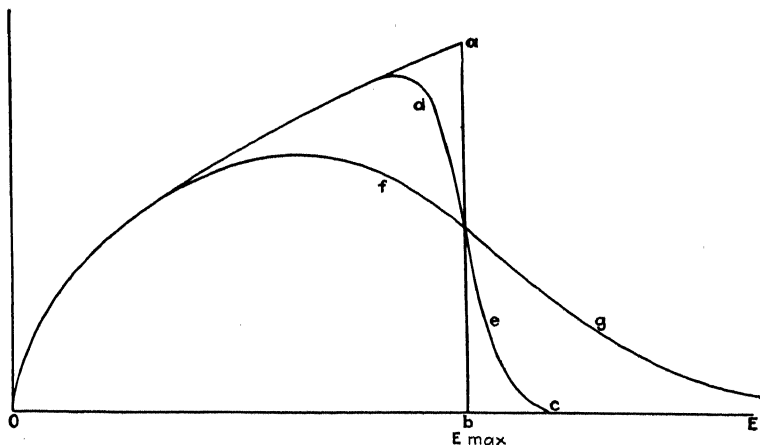


FIG. 82.—Fermi distribution in energy. Curve $oabc$ is the energy distribution at absolute zero, the other curves $odec$ and ofg being for higher temperatures.

E_{\max} , as in $odec$. A still higher temperature would be represented by ofg , and at sufficiently high temperature so that almost all the electrons were excited, the distribution would finally approach a Maxwell distribution law, the exclusion principle no longer being of importance when energies are so great that there is small chance of finding more than one electron in a state anyway. This situation would never be reached, however, in an actual metal, on account of the enormous temperature required.

The change in distribution of the electrons is not great enough to make a very important effect in conductivity. It is of great importance in thermal conductivity, on the other hand, but the place where it is most directly observed is in thermionic emission. It has been observed experimentally that hot metals

emit electrons, the number increasing enormously with temperature, as if there were a factor $e^{-(\text{energy}/kT)}$ in the law giving emission as function of temperature, where the energy must be taken to be of the order of magnitude of several volts, and is called the work function. This is explained as follows. A simple model of a metal is shown in Fig. 83, where we plot the potential in which the electrons may be assumed to move. This is constant, equal to A , throughout the metal, and the horizontal lines symbolize the Fermi levels, filled with electrons at the absolute zero. At the boundary, however, the potential must rise to a value C , higher than the maximum kinetic energy of the electrons at absolute zero, since it is observed that the electrons do not escape. This is symbolized by the rise AC , and



FIG. 83.—Potential at surface of metal. The shaded lines between A and B represent the energy levels filled with electrons at absolute zero, according to the Fermi distribution. The energy BC is the work function.

outside the metal the potential retains the constant value C . If now the temperature is raised sufficiently high so that there is an appreciable number of electrons of energy C or greater, these electrons can escape, forming those observed in thermionic emission. The number will evidently increase very rapidly with temperature, since those which can escape come from the parts e and g of the curves of Fig. 82, which increase greatly with temperature.

297. The Fermi Distribution.—It is not hard to derive the equation of the curves of Fig. 82, giving the distribution in energy in Fermi statistics. There are many ways to do it, but perhaps the simplest is by a reversal of the argument of Chap. XXXII, in which we derived the Planck black-body radiation law from Einstein's probabilities of transition. Let us first review that argument, in the reverse direction, and then see how it is applied in the present case. In Chap. XXXII we

assumed the Maxwell-Boltzmann distribution, and Einstein's probabilities, and derived from them the Planck law. We could equally well have assumed the Planck law, however, and have derived the distribution of molecular velocities from this, and the corresponding method will work for the Fermi as well as the Boltzmann distribution. Let us assume then that the density in black-body radiation of frequency ν is

$$\rho_\nu = \frac{8\pi h \nu^3}{c^3} \frac{1}{e^{h\nu/kT} - 1}. \quad (5)$$

Further, we assume that the probability of transition from a state of energy E_1 to one of energy E_2 , by absorption of radiation of frequency ν , where $E_2 - E_1 = h\nu$, is $B\rho_\nu$, and the probability of emission of radiation and of transition from the state 2 to 1 is $(A + B\rho_\nu)$, where A measures spontaneous emission, the other term forced emission, and where $A/B = 8\pi h \nu^3/c^3$. Then, if N_1 are in the state 1, N_2 in the state 2, the number of transitions from 1 to 2 per second, under the action of the radiation, will be $N_1 B\rho$, and from 2 to 1 will be $N_2(A + B\rho)$. These must be equal in thermal equilibrium, so that the net number of systems in the various states will not change with time. Hence we have

$$\frac{N_2}{N_1} = \frac{B\rho}{A + B\rho} = e^{-h\nu/kT}, \quad (6)$$

using the relations between A , B , and ρ . Thus we have

$$N_2:N_1 = e^{-E_2/kT}:e^{-E_1/kT}, \quad (7)$$

which is the Maxwell-Boltzmann distribution.

Let us now ask in what way our situation with the Fermi statistics is different. In the first place, we have a continuous distribution of energy, rather than discrete states. This is taken account of by taking an energy range dE instead of the state of energy E_1 , and another energy range dE' instead of the second state. Let the number of electrons with energy in dE be $f(E)dE$, and in dE' be $f(E')dE'$. Further, let the maximum possible number of electrons in dE be $F(E)dE$, and in dE' be $F(E')dE'$. Here F represents evidently the total number of stationary states in the interval (counting states of opposite spin as different), since each state can have but one electron. Now in computing the number of electrons going per second

from dE to dE' , we note that we are grouping together many possible transitions, since we have many possible pairs of stationary states. The probability of a transition in unit time from one level in dE to one in dE' is $B\rho$ (if the level in dE' is unoccupied) or 0 (if the level is occupied). It is in this differentiation between occupied and unoccupied levels that the Fermi statistics enter. But now we have $f(E)dE$ levels in dE , so that the probability of a transition in unit time from any one of these to one particular unoccupied level in dE' is the sum of the separate probabilities, or is $B\rho f(E)dE$. Finally, there are $[F(E')dE' - f(E')dE']$ unoccupied levels in dE' , and the chance of having a transition to any one of these is the sum of the chances to the individual ones, or $B\rho f(E)dE[F(E') - f(E')]dE'$. Similarly the probability of a transition from dE' to dE per unit time is $(A + B\rho)f(E')dE'[F(E) - f(E)]dE$. Equating these, and using the same relations before for A and B , we have

$$f(E)[F(E') - f(E')]e^{-E'/kT} = f(E')[F(E) - f(E)]e^{-E/kT}. \quad (8)$$

Divide through by $f(E)f(E')$. Then we have

$$\left[\frac{F(E')}{f(E')} - 1 \right] e^{-E'/kT} = \left[\frac{F(E)}{f(E)} - 1 \right] e^{-E/kT} = \text{constant} = A.$$

Solving for f , then, we have

$$f(E) = \frac{F(E)}{Ae^{E/kT} + 1}. \quad (9)$$

This is the distribution function for the Fermi distribution, where A is a constant independent of E (but which may depend on temperature).

The general properties of the distribution can be seen at once from the distribution function. First, at low temperatures, let us write A in the form $e^{-E_0/kT}$. Then we have

$$f(E) = \frac{F(E)}{e^{(E-E_0)/kT} + 1}. \quad (10)$$

For very small T , the exponential is e raised to a very large quantity. If $E - E_0$ is positive, this quantity is e to a large positive power, or an enormous value, so that $f(E)$ is practically zero. On the other hand, if $E - E_0$ is negative, it is e to a large negative power, or practically zero, and can be neglected compared with 1, leaving $f(E) = F(E)$. In other words, E_0 is our value E_{\max} , and as in Fig. 82 the distribution is $F(E)$ for

energies less than E_{\max} , zero for greater energies. On the other hand, for extremely high temperatures, the particles become distributed through a wide range of high energies, so that they are so spread out that the chance of a given level being filled is small. Hence $f(E) \ll F(E)$, or $Ae^{E/kT} \gg 1$. Then the distribution approaches $f(E) = \text{constant } F(E)e^{-E/kT}$, or the Maxwell-Boltzmann law. At intermediate temperatures, however, such as those concerned in thermionic emission, we are interested in the distribution for energies large compared with kT . Here we have $(E - E_0)$ large compared with kT , so that $e^{(E-E_0)/kT}$ is large compared with unity, and we can write

$$f(E) = F(E)e^{-(E-E_0)/kT},$$

again the Maxwell-Boltzmann law. This is the form of distribution which, as we stated above, is indicated by the observations on thermionic emission, and in particular the energy $E - E_0$, where E is the minimum energy necessary to escape from the metal, or BC in Fig. 83, is the work function. We see, therefore, that this is the work required to remove the most loosely bound electron of the Fermi distribution from the metal. This is found experimentally to be of the order of magnitude of two or three volt-electrons, so that in Fig. 83 the energies AB and BC are of the same order of magnitude, but AB (being about 10 volts or more) is the larger.

Problems

1. Taking the dimensions of a copper crystal lattice, and assuming one conduction electron per atom, apply the Fermi method to find the maximum kinetic energy, in volts, of the electrons at absolute zero.

2. To compress a metal at the absolute zero, we must squeeze the electrons into smaller volume, therefore increase the kinetic energy, and do work. This accounts for the larger part of the resistance to larger compressions. Find the formula for pressure as a function of volume, assuming only this repulsive effect.

3. Show that the formula for pressure as function of volume, found in Prob. 2, is the same that one would get by ordinary gas theory for the adiabatic compression of a gas with the same kinetic energy as the electron gas.

4. A simple model of a metal may be made by assuming the repulsion of Prob. 2, and an ionic attraction, giving a potential inversely proportional to the grating space. Using such a potential, determine its arbitrary coefficient by making the grating space agree with the observed value, and compute the compressibility in terms of the constants of the system.

5. Apply the method of Prob. 4 to the case of copper. Compare the resulting compressibility with experiment.

6. At a distance 0.04 atomic units ($= 0.53 \times 10^{-8}$ cm.) from the nucleus, in a rubidium atom, the potential energy of an electron is about -1540 atomic units ($= 13.54$ volt-electrons), and the density of charge is given by saying that the number of electrons per unit increase of r (measuring again in atomic units) is 42. Find how nearly this agrees with the maximum density allowed by the Fermi-Thomas method, if the maximum total energy of the electrons is zero.

7. Find the distribution function for the number of electrons whose x component of momentum is between p_x and $p_x + dp_x$, at any temperature, using Fermi statistics.

8. Of all the electrons striking the surface of a metal, only those whose momentum normal to the surface is connected with a term in the kinetic energy greater than the work function can escape. Derive the expression for the number per second escaping. In doing this, note that the number striking 1 sq. cm. of the wall per second is the number contained in a cylinder of base 1 sq. cm., slant height the vector velocity of the electron.

8. Show that if atoms and molecules obeyed Fermi statistics, the maximum kinetic energy at absolute zero, and consequent departure from the Maxwell-Boltzmann law at higher temperatures, would be so small that they would not ordinarily be observed. Actually molecules do not obey Fermi statistics, but another sort, called Bose statistics, which involves deviations from the Maxwell-Boltzmann law of about the same amount, though in the opposite direction, resulting in a reduction rather than an increase of mean kinetic energy and gas pressure.

CHAPTER XLII

DISPERSION, DIELECTRICS, AND MAGNETISM

The most important properties of atoms are connected with their interaction with each other, to form molecules and solids and all sorts of material systems. But another important set of properties is connected with the behavior of substances in external fields, electric and magnetic. We shall consider these in the present chapter. First is the question of electric fields, both varying and constant. A periodically fluctuating electric field is the same thing as a light wave, and the first problem is dispersion, the question of the forced dipole set up in an atom by the light wave. We shall find that this has just the form of the forced motion of a linear oscillator under an external sinusoidal field, which we have used earlier as a model for the electrons in dielectrics and transparent media. This agreement of form between the displacement of a forced dipole and of an atom on quantum theory is the basis of the whole classical theory of dielectrics and dispersion. Since the individual atoms behave in the manner we have already assumed, we shall not have to repeat the earlier analysis of the reaction of the dipoles back on the field, and their effect in determining the index of refraction. From dispersion we can at once pass to the behavior of dielectrics, remembering that the polarizability and dielectric constant are derived as limiting cases of dispersion for zero frequency. In connection with the polarizability, we can verify the results mentioned in Chap. XXXV about polarization and Van der Waals' forces between atoms, questions answered by similar mathematical methods. We then pass on to a general discussion of dielectrics. These show dielectric properties for two reasons: because the individual atoms show polarization, the effect mentioned before, and because molecules can possess permanent dipole moments, which become oriented under the action of an external electric field, producing a polarization and a contribution to the dielectric constant. This second type of dielectric action decreases as the temperature increases,

since temperature agitation tends to prevent the necessary orientation of the dipoles. Finally we shall pass briefly to some magnetic properties of substances. Paramagnetic substances have atoms and molecules with permanent magnetic dipoles, just like the permanent electric dipoles mentioned above, and under the action of a magnetic field these orient, so that the theory we develop for dielectrics can be used without change for paramagnetism.

298. Dispersion and Dispersion Electrons.—In Chap. XXIV, we have seen that an electron of charge e , pulled to a position of equilibrium by a linear restoring force such that its natural frequency is ν_0 , is set into vibration by an electric field $E \cos 2\pi\nu t$ in the x direction. If x is the resulting displacement of the electron, ex will be its electric moment, and this is given by

$$ex = \frac{e^2}{4\pi^2 m} \frac{E \cos 2\pi\nu t}{\nu_0^2 - \nu^2}. \quad (1)$$

The quantity $\alpha = \frac{e^2}{4\pi^2 m} \frac{1}{\nu_0^2}$ by which the field must be multiplied to get the electric moment, in the special case of zero frequency, is the polarizability of the vibrator. Now we have seen that the contribution of the induced dipoles to the external field results in a changed velocity of propagation of the wave, and hence a refractive index. We need not go through the argument again.

If now an atom has several types of vibrating electrons, rather than one type only, we must add the electric moments due to each. Thus if there are f_1 electrons in the atom with natural frequency ν_1 , f_2 of frequency ν_2 , etc., the total moment is

$$ex = \frac{e^2}{4\pi^2 m} \sum_i \frac{f_i}{\nu_i^2 - \nu^2} E \cos 2\pi\nu t. \quad (2)$$

Experimentally, such a formula gives a good value of the index of refraction as function of frequency, except for the fact that we have neglected damping, so that this formula goes to infinity at each natural frequency, instead of merely going to large values. To get agreement with experiment, we must assume that the various natural frequencies ν_i are the frequencies of light which the atom can absorb in going from the state it is in (usually the normal state) to some other; that is, they are the

frequencies determined by the quantum theory, including not merely the optical absorption frequencies in which a loosely bound electron becomes excited, but also the x-ray absorption frequencies. For instance, a sodium atom has absorption at the various discrete frequencies connected with absorption of the lines of the principal series (its 3s valence electron being excited to a p state); it has continuous absorption beyond the limit of this series, in the ultra-violet, the 3s electron being entirely removed. But also it has absorption if one of the 2 quantum electrons is ejected (the L absorption edge), or if a 1-quantum electron is ejected (K absorption). These latter are in the x-ray region. To make the classical formula fit the observations, we must assume that all these frequencies, some discrete and some distributed continuously, are connected with oscillators of suitable frequencies ν_1 .

With this interpretation of the frequencies ν_i , it is obvious that the numbers f_i of dispersion electrons of the various frequencies cannot all be integers. For there are infinitely many lines, and yet but a finite number of electrons in the atom. As a matter of fact, in the principal series of sodium, the experimental f values associated with the various lines have been determined. For the first line (the well-known D line), the f is almost, but not quite, unity; but the other lines of the series are much weaker, and decrease rapidly in dispersive power as we go down the series, so that the sum of all other f 's, for both the discrete members of the series and the continuous absorption at the end, is only a few per cent of unity. The x-ray dispersion terms correspond to fairly large f values, though on account of the factor ν^2 in the denominator, the effect on the refractive index is very small in the x-ray region. As a matter of fact, the sum of f values for the L absorptions is of the order of magnitude of the number of electrons in the L shell, and for the K absorptions of the order of magnitude of the number in the K shell. The total sum of all f 's for the atom, then, is of the order of magnitude of the number of electrons in the atom. As a matter of fact, the sum of f 's proves to be exactly equal to the number of electrons. This was first found experimentally in the following way: the forced dipole moment determines scattering as well as dispersion, as we have seen in Chap. XXV, and at frequency large compared with any natural frequency,

the displacement ex is equal to $\frac{e^2}{4\pi^2m} \frac{1}{-\nu^2} \sum_i f_i E \cos 2\pi\nu t$, leading

to the Thomson scattering formula for x-rays. Now experimentally the Thomson formula is found to hold, if we put in the actual number of electrons in place of $\sum_i f_i$; as a matter

of fact, this experiment was one of the earliest ways of determining the number of electrons in an atom. That number, then, must equal the sum of all f values. The result can now be proved theoretically as well; we shall shortly derive values for the f 's in terms of quantum theory, and the theorem can be proved by quantum theory from these values.

299. Quantum Theory of Dispersion.—We shall now derive from wave mechanics the forced dipole moment set up by a vibrating field, and show that it has the same form as the classical value. The first problem is to consider the perturbation of the atomic wave function produced by the external field. This has already been investigated in Chap. XXXII, in connection with the absorption of radiation, which, of course, is intimately connected with the dispersion. In that chapter the following problem was solved: assuming that u_m^0 was the wave function, E_m^0 the energy level, of a problem with the Hamiltonian H^0 , we added an external field, of which we here take the term of one frequency only, so that the whole Hamiltonian is $H = H^0 - exE_\nu \cos 2\pi\nu t$ (where here the phase factor in the external field is neglected).

We showed that the perturbed wave function was $\psi = \sum_m C_m(t)$

$u_m^0(x)$, where $C_m = c_m(t)e^{-\frac{2\pi i}{h}E_m^0 t}$, and where expressions were derived for the c 's as functions of time [see Eq. (16), Chap. XXXII]. Rather than writing these formulas in the general case, we make several specializations. First, it will be assumed that the atoms are practically all in the state 0, so that squares of C 's connected with other states can be neglected. Further, we neglect certain constant terms in the c 's, which do not vibrate in the same frequency as the external field, and hence cannot contribute to the dispersion. Lastly we give formulas for the C 's rather than the c 's. When this is done, we have

$$C_m = \frac{(ex)_{m0} E_\nu}{2} \left(\frac{e^{2\pi i \nu t}}{h\nu + E_m^0 - E_0^0} - \frac{e^{-2\pi i \nu t}}{h\nu - E_m^0 + E_0^0} \right) e^{-\frac{2\pi i}{h}E_0^0 t}, \quad (3)$$

where $(ex)_{m0}$ is the matrix component of the electric moment in the x direction associated with the transition from state 0 to m . Using these values of the C_m 's, and the value unity for C_0 , we can compute ψ , and from it find the mean value of the induced electric moment in the x direction, which is $\int \bar{\psi}(ex)\psi dv$. In the double sum coming from multiplying ψ by its conjugate, there are three types of terms: first, the term unity coming from the terms C_0 of both ψ and $\bar{\psi}$, which gives nothing because (ex) has no diagonal matrix component corresponding to the state 0; second, terms proportional to the C 's, coming from one factor of unity and one other C ; these are the essential terms which we consider; and third, terms in C^2 , which we neglect. In each term, we must integrate a quantity like $u_0(ex)u_m$, giving an additional factor $(ex)_{m0}$. When we carry out these integrations, and combine conjugate exponentials to get cosines, we have easily

$$\begin{aligned} ex &= \sum_m (ex)_{m0}^2 \left(\frac{1}{h\nu + E_m^0 - E_0^0} - \frac{1}{h\nu - E_m^0 + E_0^0} \right) E_\nu \cos 2\pi\nu t. \\ &= 2 \sum_m \frac{(ex)_{m0}^2 \nu_{m0}}{h} \left(\frac{1}{\nu_{m0}^2 - \nu^2} \right) E_\nu \cos 2\pi\nu t, \end{aligned} \quad (4)$$

where $h\nu_{m0} = E_m^0 - E_0^0$. This is the same as the classical formula, if we set $f_i = \frac{8\pi^2 m}{e^2 h} (ex)_i^2 \nu_i$.

It is interesting to note that this expression for the number of dispersion electrons can be written in terms of the probability coefficient B determining the probability of absorption. This, as found in Chap. XXXII, is equal to $B_{m0} = \frac{8\pi^3}{3h^2} (ex)^2_{m0}$. Thus we

have $f_{m0} = \frac{3}{\pi} \frac{m}{e^2} B_{m0} h\nu_{m0}$. It is reasonable that we should be able to do this, on account of the intimate relation between absorption and dispersion. As a matter of fact, measurements on dispersion are often used to determine the B 's, particularly in the visible range, and the converse is sometimes done as well, as in the x-ray region, where the absorption, and B , are easy to measure, but the dispersion is so small as to be difficult to get accurately.

300. Polarizability.—Setting the external frequency equal to zero, we have

$$\alpha = 2 \sum_m \frac{(ex)^2_{m0}}{h\nu_{m0}} = 2 \sum_m \frac{(ex)^2_{m0}}{E_m^0 - E_0^0}, \quad (5)$$

as the polarizability of the atom. This formula can easily be found directly by second-order perturbation theory, without going through all the calculation we have made. First we notice that the energy of a dipole of polarizability α , in a field E , is $-(\alpha/2)E^2$. To prove this, we note that it requires no work to put the unpolarized dipole in the field. When it is polarized, the internal energy of the dipole, connected with the restoring force, becomes $(\alpha/2)E^2$, or half the product of force (eE) and displacement ($\alpha E/e$). But at the same time the potential energy in the external field becomes $-\alpha E^2$, equal to the force eE times displacement $\alpha E/e$, with negative sign because the charge is being pulled in the direction of the field, and therefore to a lower potential. Adding the two, the result is $-\frac{\alpha}{2}E^2$, as we have stated.

We can now directly compute the same energy for the atom by perturbation theory, and by comparison derive the polarizability.

The perturbative energy from a field E is $-eEx$. We now have, from Chap. XXXII, the expression

$$E_n = H_{nn} - \sum_{k \neq n} \frac{H_{kn}H_{nk}}{H_{kk} - H_{nn}}$$

for the energy of the system, correct to perturbations of the second order, where H_{nk} , etc., are matrix components of the whole energy. Assuming that the unperturbed energy has a diagonal matrix, and that its value connected with the normal state (whose polarizability we compute) is E_0^0 , we have

$$E_0 = E_0^0 - \sum_{k \neq 0} \frac{H_{k0}^2}{E_k^0 - E_0^0}.$$

In this expression, we have noted that the perturbative energy has no diagonal terms, so long as the atom has no dipole moment. This is always true with the normal states of atoms. In excited states, it is in some cases possible to set up wave functions which have a dipole moment, the electrons spending more time on one side of the nucleus than on the other, and in these cases there can be a first-order effect of the electric field on the energy levels. This is observed in the spectrum, as a displacement of the lines under the action of an electric field, and is called a first-order Stark effect. Most atoms, however, show only a second-order

Stark effect, the term we have found proportional to the square of the small matrix components H_{k0} , and therefore to the square of the field. The only exceptions prove to be hydrogen, and hydrogen-like states of other atoms.

Inserting now the value of the matrix of the perturbative energy, our expression becomes

$$E_0 = E_0^0 - \sum_{k \neq 0} \frac{(ex)_{k0}^2}{E_k^0 - E_0^0} E^2 = E_0^0 - \frac{\alpha}{2} E^2, \quad (6)$$

leading to the same value of polarizability as before. It is interesting to write the polarizability in terms of the number of dispersion electrons. Then it becomes

$$\alpha = \frac{1}{m} \left(\frac{eh}{2\pi} \right)^2 \sum_k \frac{f_{k0}}{(E_k^0 - E_0^0)^2}. \quad (7)$$

In this formula, the polarizability is expressed in terms of quantities all of which can be determined from experiment. It can be used, therefore, as a direct test of the theory; or it can be used to find unknown polarizabilities from known members of dispersion electrons and energy levels, or *vice versa*.

301. Van der Waals' Force.—The Van der Waals' force which we have discussed in Chap. XXXV can be derived by second-order perturbation theory, in a way similar to that which we have used in discussing polarizability. In this case, instead of considering the perturbation of an atom by an external field, we consider its perturbation by another atom. Let us imagine two atoms at distance R , and investigate their mutual perturbations. If the first forms instantaneously a dipole of moment μ , this will produce at the second a field proportional to μ/R^3 , times a function of the angle between μ and the line joining the atoms. This will produce a perturbative energy proportional to the instantaneous dipole μ' of the second, times the field, or $\mu\mu'/R^3$ times functions of the angles of both dipoles. This then is the perturbative energy which we must use in our perturbation problem. As with the polarization by a constant field, the average value vanishes, and we must take second-order perturbations. For this, we need the nondiagonal matrix components of the quantity $\mu\mu'$, with the functions of angle. If we consider the transition from the normal state to the state in which the first atom is in

the k th state, the second the l' state, the matrix component is the product of the $0k$ component of μ , and the $0l'$ component of μ' , together with the result of integrating the functions of angle. On account of the spherical symmetry of each atom, the matrix component of μ can be determined from that for (ex) , and similarly that of μ' can be found from that for (ex') , the numerical factors coming from the angles. The final result, then, of taking the matrix component and averaging over angles is to replace the matrix component of $\mu\mu'/R^3$ by constant times $(ex)_{k0}(ex')_{l'0}/R^3$, where the constant can be determined by carrying through the computation. Next we need the diagonal elements of the total energy; these are simply the sums of the unperturbed energies of the two atoms. We finally have for the perturbed energy, then,

$$- \sum_{k,l'} \frac{\text{constant}}{R^6} \frac{(ex)_{k0}(ex')_{l'0}}{E^0_k + E^0_{l'} - E^0_0 - E^0_{0'}}.$$

When the whole computation is carried out, the constant proves to be 6.

In most cases, one transition, or one group of transitions, proves to make the major contribution to the summations for either the polarizability or Van der Waals' force. Thus for sodium we have pointed out that the D line, the first line of the principal series, has an f value far greater than the other lines; we should make small error if we left the others entirely out of account. Similarly for an inert gas, or an ion in the form of a closed shell, the first line of the absorption series has such a large frequency that it is nearly as great as the ionization potential, and all the essential terms are concentrated close to the ionization potential. In either case, then, we may approximately replace the summation by its one principal term. If we let ΔE represent the energy difference between the lowest state and this particular excited state, we have for the polarizability

$$\alpha = 2 \frac{(ex)^2}{\Delta E}.$$

In this case, the Van der Waals' potential is

$$- \frac{6}{R^3} \frac{\alpha\alpha'\Delta E\Delta E'}{4(\Delta E + \Delta E')} = - \frac{3}{2} \frac{1}{R^3} \frac{\Delta E\Delta E'}{\Delta E + \Delta E'} \alpha\alpha'. \quad (8)$$

This is a useful formula expressing the Van der Waals' potential between any pair of atoms in terms of the principal energy differ-

ences of the two (which can be estimated from the spectrum) and the polarizabilities, which can be observed or calculated. In case the two atoms are alike, this reduces to

$$-\frac{3}{4} \frac{1}{R^6} \alpha^2 \Delta E, \quad (9)$$

a useful expression which proves to be fairly accurate experimentally. We may rewrite this last formula in terms of the expression for α , obtaining

$$-\frac{3}{2} \frac{1}{R^6} \alpha (ex)^2, \quad (10)$$

a formula equivalent, to the one — constant $\frac{\alpha \mu^2}{R^6}$ of Sec. 254, Chap. XXXV.

302. Types of Dielectrics.—All atoms are polarizable, and as a result every substance acquires a dipole moment under the action of an external electric field, and shows dielectric properties. These properties can be found from the polarizabilities we have computed, the dielectric constant being given by $\epsilon = 1 + 4\pi \Sigma \alpha$, where the summation is over all atoms in a cubic centimeter, or

more accurately by $\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi}{3} \Sigma \alpha$. This dielectric constant is

obviously independent of temperature. But in addition, as we have mentioned above, there is a dielectric effect on account of orientations of permanent dipoles in the field. While atoms never have permanent dipole moments, molecules often do, as for instance NaCl, in which naturally the Na in general is positively charged, the Cl negatively. The moments tend to set themselves parallel to the field. In a solid this is hardly possible (with a few exceptions in which water molecules are free to rotate within the solid), but in liquids and gases the molecules are free to orient themselves, and a polarization results from the process. There are several features which distinguish this form of dielectric action from the other. First, it is dependent on temperature, as we shall prove, decreasing on account of temperature agitation at high temperature. Second, it shows quite a different dependence on frequency from the other sort. The molecules are fairly heavy and hard to rotate. Thus while low frequency alternating fields as well as static fields are able to orient the dipoles, higher frequency fields are too fast for them, and this type of dielectric

action drops out, usually becoming negligible in the wave length range of a few centimeters, so that for visible light there is no trace of it. An example is water, with an enormously large static dielectric constant, on account of the large moments of its molecules, but with a perfectly normal refractive index, the dipole effect being entirely ineffective at optical frequencies.

303. Theory of Dipole Orientation.—The mean electric moment resulting from dipole orientation is easily found by application of the Maxwell-Boltzmann distribution. If it were not for the field, the number of molecules with their dipole moments pointing along the directions contained within a certain solid angle would simply be proportional to the solid angle, there being no preferred direction. That is, if θ is the angle between the axis of the moment and the z axis, the number between θ and $\theta + d\theta$ would be proportional to the solid angle between these two angles, or to $2\pi \sin \theta d\theta$. If, however, there is an electric field along the z axis, there will be a potential energy depending on θ : a given molecule will have a potential energy equal to the negative of the field times the projection of the moment along the field, or $-\mu \cos \theta E$, where E is the field. Then by the Maxwell-Boltzmann law, the number having direction between θ and $\theta + d\theta$ will no longer be proportional merely to the solid angle, but to this times $e^{-(\text{energy}/kT)}$. In other words, the number between θ and $\theta + d\theta$ is proportional to

$$e^{\frac{\mu E}{kT} \cos \theta} 2\pi \sin \theta d\theta. \quad (11)$$

We may now use this distribution function to find the mean value of $\mu \cos \theta$, the component of moment along the z axis. This is evidently

$$\mu \frac{\int e^{\frac{\mu E}{kT} \cos \theta} \sin \theta \cos \theta d\theta}{\int e^{\frac{\mu E}{kT} \cos \theta} \sin \theta d\theta},$$

where the integrations are from 0 to π . To carry out the integration, let $\mu E/kT = c$, $\cos \theta = y$, so that $\sin \theta d\theta = -dy$. Then, noting that y goes from 1 to -1 , we have for the mean moment along z ,

$$\mu \frac{\int_1^{-1} e^{cy} y dy}{\int_1^{-1} e^{cy} dy} = \mu \left(\frac{e^c + e^{-c}}{e^c - e^{-c}} - \frac{1}{c} \right) = \mu \left(\coth c - \frac{1}{c} \right) \quad (12)$$

Ordinarily we are interested only in temperatures high enough, or electric fields small enough, so that c is a small number. Then approximately the mean moment is the value found by expanding our expression, which proves easily to be $\frac{1}{3} \mu c = (\mu^2/3kT) E$. This then gives a contribution to the total polarizability, and to the dielectric constant, inversely proportional to the temperature, as we have said. This temperature variation can be used experimentally to separate the two sorts of dielectric action; a plot of $(\epsilon - 1)/(\epsilon + 2)$ against $1/T$ should give a straight line, the intercept at $1/T = 0$ giving the contribution of the polarizability of the molecules, and the slope determining the permanent dipole moment. Such measurements are used to find dipole moments experimentally, and in turn this gives information about the structure of the molecules.

304. Magnetic Substances.—We shall give only a short summary of the magnetic properties of substances. In the first place, there are two fundamentally different types of magnetic behavior, as there are for dielectrics. The first is diamagnetism; any system containing electrons has induced currents set up in it when a magnetic field is impressed, and these currents in turn act like magnetic dipoles, always opposing the original field, and therefore producing a small negative contribution to the magnetic susceptibility, the magnetic analogue to polarizability, defined as magnetic moment per unit volume per unit field. The other is paramagnetism, the result of the orientation of permanent magnetic dipoles. We shall not work out the theory of diamagnetism, but shall merely state that the diamagnetic susceptibility of an atom is proportional to the mean square radius of its electrons, and measurement of this quantity is useful in checking calculations of atomic structure, since the mean square radius can be easily computed from an atomic wave function. As we have stated, all substances show diamagnetism, and only those containing permanent magnetic dipoles show paramagnetism. Where the latter occurs, however, it is almost always large enough to mask the diamagnetism, and leave a net positive susceptibility, and a magnetic permeability greater than unity.

Permanent magnetic dipoles result from two things, orbital motion, and electron spin. In single atoms, both are of importance. We have seen that the orbital angular momentum

$\frac{Lh}{2\pi}$ and spin angular momentum $\frac{Sh}{2\pi}$ of an atom unite to form the vector sum $J\frac{h}{2\pi}$. Each of these vectors carries with it a magnetic moment, but the magnetic moment associated with the orbital angular momentum proves to be $e/2mc$ times the angular momentum, while that associated with spin is e/mc times its angular momentum. Hence the total magnetic moment is not simply proportional to J . It turns out that the atom may be considered to rotate about J , and only the component of magnetic moment in this direction is of significance. This component can then be written $Jg\frac{h}{2\pi}\frac{e}{2mc}$, where g is 1 if the angular momentum arises entirely from orbital motion, 2 if it is entirely from spin, and in between for intermediate cases. Now J is quantized in space, so that its component along the preferred direction, in this case the direction of the magnetic field, is M . Hence the component of magnetic moment of the atom in the direction of the field is $Mg\frac{eh}{4\pi mc}$. It is this component which contributes to the magnetic susceptibility. It is also responsible for an observable effect in the spectrum: in the field H , the energy shift is $Mg\frac{eh}{4\pi mc}H$, and this shift in energy levels shifts and splits up the lines, removing the degeneracy in the matter of orientation. This effect of magnetic fields on spectra is the Zeeman effect, and it has been of great use in developing the theory of spectra, principally on account of the complicated but definite way in which g depends on L and S .

Since the magnetic energy of atoms depends on the orientation of J in space, there will be a preference for the atoms to line up with the field, just as there was with dipoles in the dielectric case. Again we can use the Maxwell-Boltzmann law to find the mean moment parallel to the field. Since the orientation is quantized, we should properly use the quantum statistics, but careful analysis shows that the final result comes out in important cases the same as the classical case we have already investigated. Hence we can simply use the same formula we did for dielectrics.

In the magnetic properties of solids, there are various possibilities. The general situation is that the orbital motions are so interfered with by neighboring atoms that there is no longer an

orbital angular momentum which stays constant and is quantized. The only remaining magnetic moment then arises from the spins, and this produces the paramagnetism. Thus in the iron group, the ions of the metals show large multiplicity in their spectra. As a result, they must have large spins, and these substances show paramagnetism, both when the ions are in water solution, and when they are in crystalline salts. The magnitude of the paramagnetic moment can be found from magnetic measurements, and it is found to agree with the hypothesis that it arises entirely from spins. On the other hand, in the rare earths, we recall that the group of $4f$ electrons which is being built up lies far beneath the surface of the atom. These electrons, then, are relatively unaffected by other atoms, and it appears that they have quantized orbital as well as spin angular momentum, and that this takes part in the paramagnetism.

The most spectacular magnetic property is ferromagnetism, as exhibited in iron. In this case the spins alone are oriented. But here a new feature enters, quite different from what is found in paramagnetism. It appears from experimental evidence that iron, even when not magnetized as a whole, still consists of a great many small grains, small compared with ordinary dimensions but large compared with atomic sizes, perhaps a few hundred atoms on a side, each of which acts like a permanent magnet. The process of magnetizing the iron as a whole consists of bringing the moments of these grains into parallelism with the field; for it appears that while a grain always has a magnetic moment, this moment can be rotated with respect to the grain. Magnetic saturation appears when all the grains have their moments parallel. Hysteresis is a result of the fact that the grains do not like to change their orientation, interposing a sort of friction to rotation, so that they can retain a magnetic moment even in the absence of a field. These conspicuous properties, then, are a result of the grain-like structure of the metal. One of the most convincing pieces of evidence for this structure is found in the Barkhausen effect, an effect observed when the magnetization is measured very accurately as a function of the magnetizing field, while the field is being applied. It is found that the magnetic moment increases in jumps, rather than continuously, each jump corresponding to the orientation of the moment of a whole grain.

The more difficult problem connected with ferromagnetism proves to be, not the explanation of the magnetization curve, but the question why the individual grains are permanently magnetized. Qualitatively it seems to be because the spins in adjacent atoms are coupled together, so that they wish to set themselves parallel to each other. But the magnetic forces between spins, which of course would tend to produce just such a lining up, prove to be many times too small to account for the effect. The order of magnitude of the necessary energy can be computed, and it proves to be comparable with the electrostatic binding energies observed in molecular structure. It has therefore been suggested by Heisenberg that the problem is essentially like one in molecular binding. We recall that in discussing the structure of H_2 by the method of Heitler and London we found two levels, a singlet and a triplet, the singlet lying lower because the exchange integral $(ab/H/ba)$ was negative. Heisenberg assumes that in the case of the magnetically active electrons of iron the corresponding integral is positive, so that the triplet lies below the singlet. In other words, it is more stable for the two spins to be parallel than for them to be opposite. If the same sort of thing were true all through a grain of the crystal, all the spins would tend to line up, and the energies involved would be large, comparable with binding energies, as they are experimentally. There is difficulty in supposing that the explanation is as simple as this, but it seems certain in any case that the forces making the spins of a single grain parallel to each other are in general of the nature of electrostatic, molecular interactions.

Problems

1. Compute the index of refraction of sodium gas, assuming that it has one dispersion electron connected with the D line ($5,890 \text{ \AA.}$), six connected with the L absorption edge, and two with the K absorption edge. Carry the calculation down to x-ray wave lengths, and show that the index in the x-ray region differs only slightly from unity.

2. Show that a gas consisting entirely of excited atoms shows "negative dispersion" about the possible emission lines, a contribution to the index of refraction of the opposite sign to the normal contribution. Show that this may be of importance in a real gas in that the ordinary dispersion connected with a transition up from the normal state may be diminished by excited atoms.

3. Assuming that the D line of sodium has one dispersion electron, compute the transition probabilities A and B , and find the mean life of an atom in the excited state before it radiates to the normal state.

4. The polarizability of the hydrogen atom is $4.5 a_0^3$. Using the formula for polarizability in terms of number of dispersion electrons and energy levels, and remembering that hydrogen has one dispersion electron, find the range of $E_k - E_0$ making the important contribution to the sum, and find where E_k lies in the term system.

5. Helium has a polarizability 0.20×10^{-24} , and two dispersion electrons. Show that the most important terms in its dispersion come from the continuous spectrum beyond the series limit.

6. Using the polarizability and principal energy difference from Prob. 5, find the Van der Waals' potential between two helium atoms.

7. The molecule of water has a permanent dipole moment of 1.8×10^{-18} e.s.u. Compute its dielectric constant at room temperature.

8. Using the selection rule that M changes by only ± 1 or 0, in a transition, show that a spectral line for which $g = 1$ is broken up by a magnetic field into three lines, one undisplaced, the others displaced by the frequency $eH/4\pi mc$ in either direction from the original line.

9. Compute and draw the Zeeman patterns for the sodium D lines, $^2P \rightarrow ^2S$, using the information that $g = \frac{3}{2}$ for $^2P_{\frac{3}{2}}$, $\frac{1}{2}$ for $^2P_{\frac{1}{2}}$, 2 for $^2S_{\frac{1}{2}}$.

10. Set up the problem of finding the mean component of magnetic moment of an atom along the direction of the field, using quantum statistics.

11. Using the classical formula of the derivation in the text, plot magnetic moment of a dipole in the direction of the field, for large fields, showing that it approaches a constant value, or saturation, at sufficiently large fields. Prove that this corresponds to having all dipoles oriented along the field.

SUGGESTED REFERENCES

In a single volume like the present one, it is impossible to do justice to many branches of theoretical physics, and some are hardly touched on. The student following the subject further will wish to refer to textbooks in the older parts of the field, original papers in more modern parts. The references which we give in the present section are far from a complete list, and many good books are not included, but it seems worth while to suggest a few texts to which the student who is familiar with the present book can refer, without too great difficulty.

First, there are a number of other texts which, like the present one, give a general survey of theoretical physics. Since they show a wide variety in their approach and emphasis, they are often worth consulting. Among these may be mentioned "*Introduction to Theoretical Physics*," by L. Page (Van Nostrand), which gives a good account of classical physics; "*Introduction to Mathematical Physics*," by Houstoun (Longmans), containing discussion of potential theory, hydrodynamics, electromagnetic theory, wave motion, and thermodynamics; "*Introduction to Theoretical Physics*," by W. Wilson (Methuen), of which Vol. I, covering mechanics and thermodynamics, has appeared at the date of writing this; and "*Introduction to Theoretical Physics*," by A. Haas (Van Nostrand), treating modern as well as classical theoretical physics. Two longer treatises on theoretical physics, in several volumes, may also be mentioned: "*Introduction to Theoretical Physics*," by M. Planck (Macmillan), an English translation of a well-known German text, and "*Einführung in die theoretische Physik*," by C. Schaefer (W. de Gruyter). These last two works go a good deal more into detail than is possible in the present book.

In addition to these works on general theoretical physics, the student will doubtless have occasion to consult books on mathematical analysis. A good book on advanced calculus, such for instance as "*Advanced Calculus*," by E. B. Wilson (Ginn), will be helpful. At the same time a more advanced book on analysis, such as "*Mathematical Analysis*," by Goursat and Hedrick (Ginn), or "*Partielle Differentialgleichungen der Physik*," by Riemann and Weber (Vieweg), will furnish much useful information. In these one will find treatment of a number of branches of mathematics which we have merely touched on, such as the theory of functions of a complex variable, theory of special functions, calculus of variations, etc. In addition to these works on analysis, a book on algebra, such as "*Modern Algebra*," by Bôcher (Macmillan), will be found helpful, particularly in studying the properties of determinants and linear transformations. Finally, "*A Short Table of Integrals*," by B. O. Peirce (Ginn), and "*Funktionentafeln*," by Jahnke and Emde (Teubner), will be found invaluable for detailed assistance in calculation. For definite integrals which are not given in these books, "*Tables*

des Integrales Définies," by Bierens de Hahn (Amsterdam), will be found a source of much information.

Next we come to a number of specific references on the various chapters. "*Dynamics*," by A. G. Webster (Teubner), is one of the most useful references on the material of the first part of the book. This treats the dynamics of particles, generalized coordinates, dynamics of rigid bodies, potential theory, elasticity, and hydrodynamics. "*Electric Oscillations and Electric Waves*," by G. W. Pierce (McGraw-Hill), takes up the material on oscillating electric circuits which we give in the first few chapters, and also material on Maxwell's equations and electromagnetic waves, which we treat in Chaps. XIX to XXVI. "*The Dynamical Theory of Sound*," by Lamb (Arnold), treats oscillations of particles, and vibrations of strings and membranes. "*Theory of Sound*," by Rayleigh (Macmillan), is the standard treatise on sound, and interprets the field in such a broad way that it is practically an introduction to theoretical physics in itself. The vibrations of particles, coupled systems, strings, membranes, and vibrating solids, elasticity, wave motion, all are carefully treated. In the mechanics of particles and in rigid dynamics "*Elementare Mechanik*," by Hamel (Teubner), will be found a useful reference. "*Gyrostatics and Rotational Motion*," by A. Gray (Macmillan), gives useful and detailed discussion of the dynamics of rigid bodies. "*Hydromechanics*," by Ramsey (Bell), may be recommended for this subject, and also "*Physics of Solids and Fluids*," by Ewald, Pöschl, and Prandtl (Blackie). For the more mathematical side of mechanics, potential theory, vector analysis, Fourier series, etc., the following are suggested: "*Vector Analysis*," by H. B. Phillips (Wiley); "*Fourier Series and Spherical Harmonics*," by W. E. Byerly (Ginn); "*Newtonian Potential Function*," by B. O. Peirce (Ginn).

For the chapters on electrodynamics and optics, there are a number of good references in addition to the chapters from the various general texts. "*Classical Electricity and Magnetism*," by Abraham and Becker (Blackie), and "*Electricity and Magnetism*," by J. H. Jeans (Cambridge), contain detailed treatment of the electromagnetic side of the subject. "*Lehrbuch der Optik*," by Försterling (Hirzel), and "*Optik*," by M. Born (Springer), are excellent treatments of optics from the standpoint of the electromagnetic theory. "*Theory of Electrons*," by H. A. Lorentz (Teubner), contains important material on electrodynamics and the electronic structure of matter, though since it was written in the early days of electron theory, before the time of quantum theory, there are many parts of it which cannot be accepted at present. Finally, "*Physical Optics*," by R. W. Wood (Macmillan), gives an excellent treatment of the more experimental side of optics.

For the last chapters, on wave mechanics and the structure of matter, there are in the first place a number of general texts. "*Atombau und Spektrallinien*," by Sommerfeld (Vieweg), is a standard work on the older forms of quantum theory, and "*Vorlesungen über Atommechanik*," by M. Born, Vol. I (Springer), contains a rather complete mathematical development of the older quantum mechanics. "*Atoms, Molecules, and Quanta*," by Ruark and Urey (McGraw-Hill), deals with general quantum theory as well as wave mechanics. "*Quantum Mechanics*," by Condon and Morse (McGraw-Hill), treats the more elementary methods of wave mechanics, as

does "*Atombau und Spektrallinien, Ergänzungsband*," by Sommerfeld (Vieweg). "*Wave Mechanics*," by N. F. Mott (Cambridge), is a short but readable account of the elementary principles of wave mechanics, and "*Wave Mechanics, Elementary Theory*," by J. Frenkel (Oxford), the first of three projected volumes, furnishes a more detailed treatment of general principles, with particular emphasis on the statistical side of the theory. The older "*Einführung in die Wellenmechanik*," also by Frenkel, gives a general survey of the field, and includes details of some of the soluble problems. For spectroscopic purposes, in addition to the references mentioned, "*Structure of Line Spectra*," by Pauling and Goudsmit (McGraw-Hill), and "*Linien-spektren*," by F. Hund (Springer), contain good treatments. Chemical applications are not adequately dealt with in any texts at present, but "*Quantum Mechanics of Chemical Reactions*," by H. Eyring (Chemical Reviews, February, 1932), contains a survey of material on reactions. Statistical mechanics and thermodynamics are treated in many places, but perhaps the most useful one from the standpoint of the structure of matter is "*Kinetische Theorie der Wärme*," by K. F. Herzfeld (Vieweg). For quantum statistics, with particular application to the structure of metals, "*Quantenstatistik*," by L. Brillouin (Springer), is to be recommended. One can hardly pass over this subject, however, without mentioning some of the standard texts: "*Elementary Principles in Statistical Mechanics*," by Gibbs (Longmans); "*Vorlesungen über Gastheorie*," by L. Boltzmann (Barth); and several more modern works, such as "*Dynamical Theory of Gases*," by J. H. Jeans (Cambridge), and "*Statistical Mechanics*," by R. H. Fowler (Cambridge). For the theory of dielectrics, "*Polar Molecules*," by P. Debye (Chemical Catalog Co.), is to be recommended, and for both dielectric and magnetic properties, "*Theory of Electric and Magnetic Susceptibilities*," by J. H. Van Vleck (Oxford), gives a detailed and excellent discussion.

Finally, in addition to specific books, the student will find it advantageous to make liberal use of the two extensive reference works, "*Handbuch der Physik*" (Springer) and "*Handbuch der Experimentalphysik*" (Akademische Verlagsgesellschaft). Both of these sets of books provide a convenient and useful source of reference in both experimental and theoretical physics and cover practically every subject in these fields.

INDEX

A

A priori probability of a group of states, 370-371
 Absolute zero, crystals at, 472ff.
 Absorption coefficient, optical, 256
 Absorption probabilities, relation to dispersion electrons, 549
 Action, principle of least, 342
 Action variable, and contact transformation, 88
 relation to correspondence principle, 359
 Activation energy, 490
 Angle variable (*see* Action variable)
 Angles, Euler's, 100-101
 Angular momentum, in quantum mechanics, 410ff.
 of rotating rigid body, 92ff.
 Angular rotation, lack of vector character, 100
 Anharmonic oscillator, 38
 Anomalous dispersion (*see* Dispersion)
 Antisymmetry of electronic wave function, 504ff.
 Aphelion, 62
 Approximate solution for non-uniform string, 147-148
 (*See also* Wentzel-Kramers-Brillouin method)
 Archimedes' principle, 196
 Artificial electric line, analogy to weighted string, 132
 Associated Legendre polynomials, or associated spherical harmonics, 171, 409
 Atom, general discussion, 406-437
 model of Fermi and Thomas, 535-536

Atom, perturbation theory applied to multiplets, 390, 518-522
 repulsion and attraction between, 439-452
 Atomic refractivity, 284
 Atwood's machine, 76
 Auger effect, 494
 Average values, of functions of coordinates and momenta, on wave mechanics, 375, 380
 in phase space, 372
 Axis of rotation, instantaneous, 99
 Azimuthal quantum number, 410, 519

B

Band spectra, 480, 483
 Barkhausen effect, 557
 Beats, with coupled oscillators, 110
 between transient and steady motion, 34
 Bent beam, 184
 Bernoulli's equation, 191-192
 Bessel's equation and function, 18, 166, 169, 170
 Bimolecular reactions, 489-491
 Binomial theorem, 4
 Biot-Savart law, 233
 Black-body radiation, Planck's law, 395
 Body forces, in elasticity, 173
 in hydrodynamics, 191
 Bohr, correspondence principle, 361
 frequency condition, 360
 quantum rules, 331
 theory of hydrogen, 411
 Boltzmann distribution law, 369
 Bose statistics, 544
 Boundary conditions, circular membrane, 166

- Boundary conditions, electromag-
netic field, 258-259
heat flow, 205
rectangular membrane, 161
string, 122
wave mechanics, one-dimensional
motion, 350-351
- C
- Canonical ensemble, 368-369, 456-
458
in quantum theory, 467
Capacity of parallel plate condenser,
215
Center of gravity, separation of
coordinates for diatomic mole-
cule, 481
Central field, 61-63
electron motion in, 418ff.
phase space for, 82
Chandler period, 99
Characteristic functions and num-
bers, 139, 151
(*See also* Wave functions)
Charge density, 211-212
Chemical compounds, 447-451
Chemical reactions, 488ff.
Circuit, electric (*see* Electric circuit)
Circular membrane, 164-168
Circular polarization, 266
Classical statistical mechanics in
phase space, 364ff.
Coefficients, of Fourier series, 124
of viscosity, 193
Coherence of light, 295-299
Collisions, quantum-theory treat-
ment, 402-404, 488ff.
Combination tones, 38
Commutation rule, 380
Complex exponentials and complex
numbers, 22-26
Compounds, chemical, 447-451
Compressibility, of crystal lattice,
471ff.
of elastic solid, 183-184
Condenser, energy in, 246-247
theory of parallel plate, 214
Conduction of electricity in metals,
electron theory, 281
quantum theory, 449, 495-496,
536ff.
Conductivity, specific electrical, 241
thermal, 197
Conductor of electricity, as an
equipotential, 215
Configuration, atomic, 430
Configuration space, 83-84, 117
Conjugate foci, 344
Conservative system, condition for,
55
Constants of integration, 11
Constraints, 75-76
Contact transformation, 87-90
and correspondence principle, 359-
361
Continuity, equation of, 186-187
for electric flux, 237
for flow of electricity, 231
relation to divergence of elec-
tric field, 211
Continuity conditions, electromag-
netic wave, 258-259
vibrating string, 156-157
Continuous medium, 120
Convection current, 238
Convergence, 5-8
of Fourier series, 125
Coordinates, curvilinear, curl in, 229
gradient, divergence, Laplacian
in, 199-201
generalized, equations of motion
in, 59ff., 69ff.
Cornu's spiral, 320ff.
Corpuscular theory of light, 329
Correspondence principle, 359-363
and quantum statistics, 466
Coulomb's law, 210
Coupled systems, 107ff.
application to radioactivity and
collisions, 402-404
Cross section of atoms, collision, 404
Crystals, 472ff.
valence binding and, 449
Curl, in curvilinear coordinates, 229
of electric field, 211
of a vector, 55

Current, density of, 231
displacement, 236-239
Curvilinear coordinates (*see* Coordinates, curvilinear)

D

D'Alembert's equation, 243
solution of, 304
Damping, critical, 28
logarithmic, 28
of vibrating string, 142-143
Davisson and Germer, experiment of, 336
De Broglie, 336
Debye's theory of specific heats, 478
Decrement, logarithmic, 28
Deformation of elastic solid, 180
Degeneracy, gas, 469-470
in multiply periodic motion, 363-364
orbital, in atomic multiplets, 521-522
perturbation theory for, 390-391
spin, 510, 521-522
in square membrane, 168-170
Density, charge, 211-212
current, 231
of energy in electromagnetic field, 249-250
Derivative, directional and partial, 54
Determinant, expansion of, 387
Determinant form of electronic wave functions, 504
Determinantal equation for string, 155
Diamagnetism, 555
Diatomic molecules, electronic energy, 522-523
nuclear motion, 481-485
Dielectric, force in an inhomogeneous, 257
Dielectric constant, 239
relation to polarization, 275
temperature dependence, 555
Dielectrics, types of, 553-554
Difference equations, 129
Differential equations, general properties, 10*ff.*
linear, properties, 36
solution by Green's method, 222-223
Diffraction, 311-328
Dipole, and dipole moment, 221
oscillating, 288*ff.*
Dipole orientation, temperature dependence, 554-555
Direction cosines, 49
Directional derivative, 54
Discontinuities in functions, Fourier representation, 126
Discontinuity, in electric field, 214, 222
in electromagnetic field, boundary conditions, 258-259
in electrostatic potential, 222
Dispersion, of electromagnetic waves in metals, 256
electron theory, 270*ff.*
quantum theory, 546-549
Displacement, electric, 239, 274
Displacement current, 236-239
Dissipation function, 142-143
Dissipative forces, 40
Divergence, in curvilinear coordinates, 200-201
of electric field, 210-211
of a vector, 56
physical meaning of, 187
Double Fourier series, 162
Double layer, 221
Double pendulum, 119
Doubly periodic motion, 63-64, 84-86
in quantum theory, 362-364
Dulong and Petit's law, 474
Dynamic stability, 103

E

Effective nuclear charge, 425, 431-432
Einstein, formula for specific heat of solids, 477
photoelectric law, 330

- Einstein, probability coefficients for radiation, 393-399
- relation between energy and mass, 252
- Elastic constants, 181
- Elastic electronic collisions, 402-403
- Elastic solid, 172-183
- Elastic waves, 179-180
- Electric circuit, with inductance and resistance, 16
- oscillations, 20, 28*ff.*
- Electric conductivity (*see* Conduction of electricity in metals)
- Electric displacement, 239, 274
- Electric field, 210-214
 - in spherical cavity, 279
- Electric moment of an atom, matrix component, 376
- Electromagnetic field, energy in, 246*ff.*
 - of oscillating dipole, 290-291
 - quantization of, 399
 - wave equation for, 243-244, 253
- Electromagnetic induction, 235-236
- Electromagnetic units, 227
- Electromagnetic waves, in metals, 256
 - polarization of, 262
 - reflection and refraction of, 258*ff.*
 - spherical, 286*ff.*
- Electromotive force, 212, 250-251
- Electron, radius of, 252
- Electron collisions, 402-404
- Electron emission from metals, 352-353, 539-540
- Electron energy, atoms, 435-437, 518-522
 - metals, 494-497, 536*ff.*
 - molecules, 483, 522-530
- Electron equivalence and exclusion principle, 519-520
- Electron excitation, collisions of atoms with, 491*ff.*
- Electron interactions, 501*ff.*
- Electron pair valence bond, 443-444
- Electron shells in atoms, 425
- Electron spin, 443, 507*ff.*
- Electron theory and dispersion, 270*ff.*
- Electron wave functions, determinant form and antisymmetry, 504-505
- Electrostatic field, energy in, 247
- Electrostatic potential, 211-212, 222
- Electrostatic problems, and potential theory, 210-217
- Ellipsoid of inertia, 95-96
- Elliptical polarization, 266
- Emission, thermionic, 352-353, 539-540
- Energy, of activation, 490
 - of atoms, 435-437, 518-522
 - electrical, 246-256
 - internal, 460
 - of ionic crystals, 473
 - mechanical, 39-46, 52
 - of metals, 494-497, 536*ff.*
 - of molecules, 483, 522-530
- Energy density of radiation, mean value, 393
- Energy levels, Bohr's hypothesis, 331
- Energy surface in phase space, 80
- Ensembles, canonical, 368-369, 456-458
 - in statistical mechanics, 365-369
- Entropy, 460
- Equation, of continuity (*see* Continuity, equation of)
 - of motion, of elastic solid, 175-176
 - of fluid, ideal, 190-191
 - viscous, 194
 - mechanical, generalized coordinates, 59*ff.*, 69*ff.*
 - of membrane, 160
 - of rigid body, 96*ff.*
 - of string, 120-121
 - in normal coordinates, 140
 - variable, 146-147
 - of state of gases, 454-470
 - of solids, 478-480
- Equations, difference, 129
- Equilibrium, stable, 45
- Equinoxes, precession of, 105
- Equipotential surfaces, 54, 215
- Equivalence of electrons, and exclusion principle, 519-520
- Ergodic motion, 81

- Euler, angles, 100–101
 equations of hydrodynamics, 190–191
 equations for rigid body, 98
 period, 99
- Even functions, 126
- Exclusion principle, free electrons, 531–533
 general discussion, 502–516
 periodic table, 426
 valence attraction, 443–444
- Expansion, Fourier, 123–128
 in normal functions for variable string, 153
 Taylor's, 4–5
 in wave functions, 382–383
- Exponential, complex, 22
- Exponential integral function, 18
- Exponential solution, vibrating particle, 21
 vibrating rectangular membrane, 161
 vibrating string, 121
- External forces, on coupled oscillators, 113
 generalized coordinates, 69–70
 motion under, 35–36
- External radiation field, perturbation of atoms by, 392–393
- F
- Falling body, 11*ff.*
- Faraday's induction law, 239
- Fermat's principle, 339–342
- Fermi statistics, 531*ff.*
- Fermi-Thomas atomic model, 535–536
- Ferromagnetism, 557–558
- Field, central (*see* Central field)
 electric (*see* Electric field)
 electromagnetic (*see* Electromagnetic field)
 electrostatic (*see* Electrostatic field)
 vector, 51
- First law of thermodynamics, 460
- Flow, of fluids, 185*ff.*
 of heat, 197*ff.*
- Flow, lines of, 186
- Fluids, flow of, 185*ff.*
- Flux, 185
 magnetic, 235
- Flux density, 185–186
 of heat flow, 198
- Force, on charge and current, 240
 external (*see* External forces)
 interatomic, 439*ff.*
- Forced vibrations, of particle, 29
 of string, 142–143
- Fourier series, 120–128
 double, 162
 in function space, 139
 generalization for multiplyperiodic motion, 362
- Fourier's method for transient heat flow, 203–205
- Fraunhofer diffraction, 315*ff.*
- Free electrons in quantum theory, 531–535
- Free energy, 458*ff.*
 of crystals, 476
- Fresnel diffraction, 315*ff.*
- Fresnel equations for reflection, 262–264
- Fresnel integrals, 320, 328
- Fresnel zones, 308*ff.*
- Function space, 137*ff.*
- Functions, odd and even, 126
 representation by power series, 2*ff.*
 scalar product in function space, 153
- f*-values for dispersion electrons, 547–548
- G
- Γ space, 365
- Gases, dispersion in, 275–278
 equation of state and general properties, 454*ff.*
- Gauss error curve, 372
- Gauss's theorem, 187–188
- General solution of differential equation, 15
- Generalized coordinates, curl in, 229
 equations of motion in, 59*ff.*, 69*ff.*
 gradient, divergence, Laplacian in, 199–201

Generalized force, 69-70
 in vibrating string problem, 141
 Generalized momentum, 61, 69*ff.*
 Geophysical problems with elastic waves, 179-180
 Gibbs-Helmholtz equation, 460
 Gradient, in curvilinear coordinates, 200
 of a scalar, 54, 56
 Green's distribution, 221-222
 Green's method for differential equations, 222-223
 Green's theorem, 217

H

Half-breadth of resonance band, 37
 Hamiltonian function, 71-72
 Hamilton's equations of motion, 71-76
 Hamilton's principle, 342
 Hartree's method of self-consistent fields, 430-431
 Heat flow, 197-209
 Heitler-London method for molecular energy, 523-527
 Hertz, electric waves, 238
 vector, 291*ff.*
 Homogeneous quadratic functions, 73-74
 Homopolar valence, 443, 523-527
 Hooke's law, 177-182
 modified for viscous fluids, 193
 Huygens' principle, 302-311
 Hydrogen atom, 406*ff.*
 Hydrogen molecule, 523-527
 Hydrostatic pressure, 174

I

Ideal fluid, 190
 Images, method of, 215-216
 Impedance, 33
 Imperfect gases, 462*ff.*
 Index of refraction, 253, 270-283
 (*see also* Dispersion)
 Induced emission, probability of, 394
 Induction, electromagnetic, 235-236
 electrostatic, 215
 Induction vector, magnetic, 239

Inelastic electronic collisions, 403
 Inertia, moment and products of, 95-96
 Infinite series (*see* Series)
 Initial conditions, for circular membrane, 168
 for quantum-mechanical motion, 377*ff.*
 for rectangular membrane, 162
 for string, 122
 for transient vibrations of particle, 29
 Inner shielding, 437
 Instantaneous axis, 99
 Integral, line, 52-53
 phase or quantum, 358-359
 of state, in kinetic theory, 458-459
 Integral method for transient heat flow, 205-208
 Intensity, of electric field, 210
 of magnetic field, 225
 of radiation, and correspondence principle, 363
 and selection principles, 417
 Interaction, of electrons, 501*ff.*
 of nuclei, perturbation method, 497-499
 Interatomic forces and molecular structure, 439*ff.*
 Internal energy of a system, 460
 Ionic compounds, 449-451
 Ionic crystals, 473
 Ionic forces, 439
 Ionization potentials, 435-437
 Iron group of elements, 430
 paramagnetism of, 557
 Irreversibility of heat flow, 208
 Irrotational flow, 188-192
 Iso-electronic sequences, 438
 Isotopes, 407

K

Kinetic energy, 41
 of rigid bodies, 95

L

Lagrange's equations, 58-67, 75
 for weighted string, 131

- Lagrangian function, 59
 with magnetic field, 77
 in relativity, 78
 Laplace's equation, for electrostatics, 212
 for heat flow, 198
 solution as surface integral, 220
 for velocity potential, 190
 for vibrating membrane, 164
 Laplacian, 56
 in curvilinear coordinates, 201
 in polar coordinates, 164-165
 Larmor precession, 106
 Legendre polynomials, 158
 associated, 171, 409
 Legendre's equation, 171
 Lenz's law, 235
 Level, energy, 331
 Lewis, electron pair bond, 443
 Line integrals, 52-53
 Linear differential equation, properties, 36
 Linear oscillator, phase space, 81
 Linear polarization, 266
 Linear restoring force, 19
 Linear transformation, 115
 Lines of flow, 186-188
 Liouville's theorem, 365-366
 Liquids, dispersion in, 278-280
 flow of, 185-195
 gases, and solids, comparison, 454
 Lissajous figures, 85
 Longitudinal waves in elastic solid, 178-179
 Lorentz force, 240
 Lorenz-Lorentz law, 280
 L-S coupling, 519
- M
- Magnetic properties, quantum theory, 555-558
 Magnetic quantum number, 410
 Magnetism, 225-234
 Many-body problem in wave mechanics, 432-433
 Matrices in quantum mechanics, 374-384
- Maxwell-Boltzmann distribution law, 369
 Maxwell's distribution of velocities, 372
 Maxwell's equations, 235ff.
 Mean values (*see* Average values)
 Mechanical energy, 39-46, 52
 Mechanics, statistical, 364-371, 454-470
 wave nature of, 335-336
 Membrane, vibrations of, 160ff.
 Metals, classical theory, electron theory, 280-283
 plane waves of light in, 255-256
 reflection of light from, 267-268
 quantum theory, electrons in, 531-543
 nature of conduction process, 494-497
 relation to molecular orbitals, 529-530
 relation to valence compounds, 449
 Method of images, 215-216
 Microcanonical ensemble, 367-368
 and quantum statistics, 466-467
 Molecular refractivity, 284
 Molecules, electronic energy, 518, 522-530
 general structure and interatomic forces, 439-451
 nuclear motions, 480-486
 Moment of inertia, 95
 Momentum, angular (*see* Angular momentum)
 generalized, 61, 69ff.
 Momentum operator in wave mechanics, 380
 Momentum space, 83-84
 Morse potential curve for molecules, 445-446
 Moseley's law, 437
 Motion of rigid bodies, 92ff.
 in several dimensions, 46
 of several particles, 117
 μ -space, 364
 Multiplets, 390, 509-527
 Multiply periodic motion, 63-64, 84-86

Multiply periodic motion, in quantum theory, 362-364
 Multivalued potential, 228
 Mutual induction, coefficient of, 245

N

Negative dispersion, 558
 Neutrons, 407
 Newton's law of motion, 11, 70-71
 Nodal line, Euler's angles, 101
 Nodes in vibrating membrane, 162, 167
 Non-central two-dimensional motion, 83
 Nonconservative systems, 41
 Nonuniform string, 146-158
 Normal coordinates, 107-114
 general theory, 134-142
 thermal vibrations of crystals, 474-475
 Normal dispersion, 277
 Normal functions, for vibrating string, 139, 151-153
 (*See also* Wave functions)
 Normal incidence, reflection coefficient, 260-262, 267
 Normal stresses, 174
 Normalization, coupled systems, 112, 116
 nonuniform string, 152
 quantum theory, 374-375, 382-383
 weighted string, 136
 Nuclear atom, 406-407
 Nuclear charge, effective, 425, 431-432
 Nuclear motions in molecules and solids, 471-499
 Nutation, 104

O

Odd functions, 126
 Ohm's law in differential form, 241
 One-electron energies and wave functions, 433-437
 Open circuits, 237

Operators in wave mechanics, 380-382

Optics, 258-333
 Orbital degeneracy, 416, 520-521
 Orbits, central motion, 62-63
 hydrogen, 412
 Orthogonality, Bessel's functions, 169-170
 coupled systems, 116
 nonuniform string, 151
 quantum theory, 378
 sine and cosine, 125
 weighted string, 136
 Oscillating dipole, radiation from, 288ff.
 Oscillations, of electric circuit, 20
 simple harmonic, 19ff.
 Oscillator, anharmonic, 38
 coupled, 107ff.
 linear, classical theory, 19ff.
 quantum theory, 354-357, 384
 two-dimensional, 84-86
 Oscillator strength, 280
 Outer shielding, 437
 Overtone, 120

P

Parallel plate condenser, 214-215
 Paramagnetism, 555-557
 Palladium group of elements, 430
 Partial derivative, 54
 Partial differential equation, 121
 Particular solution of differential equation, 15
 Penetrating orbits, 419-422
 Penetration of potential barriers, 351-353
 Penetration force between atoms, 442
 Penetration interaction for H_2 , 525
 Perfect gas, classical theory, 461
 quantum theory, 468ff.
 Perihelion, 62
 Periodic force in vibrating string problem, 142
 Periodic motion, 44
 multiply (*see* Multiply periodic motion)

- Periodic system of the elements, 426*ff.*
 Permeability, magnetic, 239
 Perturbation theory, nonuniform string, 154*ff.*
 quantum mechanics, 386-404
 Phase change, on reflection, waves
 on string, 158
 in total reflection, 266
 Phase integral, 88-91, 359-360
 Phase space, 79*ff.*, 358*ff.*
 for free electrons, 532
 Photoelectric emission, 330, 343, 540
 Photons, 330, 332
 Planck's constant, 330
 Planck's law of black-body radiation, 395
 Plane waves, elastic, 176-178
 optical, 253-256, 258-268
 Planetary motion, 60
 Point transformation, 87
 Poiseuille's law, 194-195
 Poisson's equation, 212, 217*ff.*
 Poisson's ratio, 182
 Polarizability of atoms and ions, 275, 439-441, 549-555
 Polarization, of dielectric, 270-275
 of light, 264-268, 295
 Pole of function, 5
 Polyatomic molecules, 485-486
 Polynomials, Legendre, 158
 Potential, electrostatic, 211-212
 magnetostatic, 225
 retarded, 303-305
 vector, 231
 velocity, 188*ff.*
 Potential barriers, penetration, 351-353
 Potential energy, 41*ff.*, 52-55
 Power series, 1-8
 Power-series solution, for differential equations, 10-20
 for secular equation of perturbation theory, 387-390
 Poynting's vector, 249-251
 Precession, 92-93, 104-106
 Predissociation, 494
 Pressure, elasticity, 175*ff.*
 hydrodynamics, 191*ff.*
 Pressure, kinetic theory, 459*ff.*
 of solid, 472
 Principal axes, coupled systems, 117
 of inertia, 96
 of stress, 175
 Principal quantum number, 410
 Principle of least action, 342
 Probability, a priori, 370-371
 Probability relations in wave mechanics, 333-337
 Products, of inertia, 95
 of vectors, 49-51
 Progressive waves, 149
 Protons, 407
- Q
- Quantum condition, 353-361
 Quantum defect, 422
 Quantum derivation of Einstein probability coefficients, 395-399
 Quantum hypothesis of Planck, 330
 Quantum number, 354, 410, 519
 Quantum statistics, 466*ff.*
 Fermi, 531*ff.*
 Quantum theory and phase space, 369-371
 Quasi-ergodic motion, 81
 quantum theory, 364
 statistical application, 367
 Quasi-stationary processes, 241
- R
- Radial motion in central field, 61-62
 Radial wave function, 409*ff.*
 Radiation, electromagnetic, 186-300
 perturbation of atoms by, 392-393
 quantization of, 399
 Radiation intensities and correspondence principle, 363
 Radioactivity, quantum theory, 356, 402
 Ramsauer effect, 404
 Rare earths, paramagnetism of, 557
 structure of, 430
 Rayleigh scattering, 294
 Reactance of electric circuit, 33
 Reactions, chemical, 488-494

- Rectangular membrane, 160–164
 - Reflection, elastic waves, 179
 - electromagnetic waves, 258–268
 - waves on strings, 156–158
 - Refraction, electromagnetic waves, 258–268
 - index of (*see* Dispersion)
 - Relaxation time, 257
 - Resistance, specific, 241
 - (*See also* Metals)
 - Resolving power, grating, 328
 - lens, 325–326
 - Resonance, 29–33
 - Resonance scattering, 295
 - Retarded potentials, 303–305
 - Rigid bodies, 92–105
 - Rolling-ball analogy, 45–46
 - Rotating system of axes, vectors in, 97–98
 - Rotation, of coordinates, 114–116
 - of diatomic molecule, 482
 - Rotator, quantum condition for, 354
 - Rydberg formula, 422
 - Rydberg number, 408
- S
- Scalar potential, 241–244
 - for oscillating dipole, 288–289
 - Scalar product, of two functions, 153
 - of two vectors, 49
 - Scalar quantities, 48
 - Scattering, of electrons, 404
 - of light, 293–299
 - Schrödinger's equation, 345–346
 - including the time, 381–382
 - many-body problems, 432–433
 - Second law of thermodynamics, 460
 - Secular equation, coupled oscillators, 108
 - perturbation theory, string, 155
 - wave mechanics, 386–390
 - Selection principles for spectra, 417
 - Self-consistent fields, 430–431
 - Separation of variables, method of, 163–165
 - Series, Fourier, 120*ff.*
 - power, 1*ff.*
 - Series spectra, 416–418
 - Several particles, general problem of motion, 117
 - Shearing stress and strain, 174, 177
 - Shells, electronic, in atoms, 425
 - Shielding of electrons in atoms, 425–437
 - Simple harmonic vibrations, 19*ff.*
 - Singularity of function, 5
 - Sleeping top, 105
 - Solenoid, energy in, 249
 - magnetic field in, 231
 - Solids, dispersion in, 278–280
 - elastic, 172–183
 - physical properties, 471–480
 - Sources and sinks, 186
 - Specific conductivity, 241
 - Specific heat, of gases, 460–461
 - of molecules, 483
 - of solids, 476–478
 - Specific resistance, 241
 - Spectra, of atoms, 407–418, 435–437, 509–520
 - of molecules, 480, 483
 - Spectral analysis of a light wave, 298–299
 - Spectral series, nomenclature, 416–418
 - Spectral terms, 331
 - Spherical electromagnetic waves, 286–299
 - Spherical harmonics, 202–203
 - associated, or associated Legendre polynomials, 171, 409
 - Spin of electron, 443, 507*ff.*
 - Spin degeneracy, 510–514
 - Spontaneous radiation, 393, 399–402
 - Square membrane, degeneracy, 170
 - Stability, dynamic, 103
 - Stable equilibrium, 45
 - Standing waves, 149
 - Stark effect, 550
 - Stationary states, 331
 - and perturbation theory, 400–402
 - Statistical interpretation of wave theory, 332–333
 - Statistical mechanics, 364–371, 454–470
 - Steady flow, of fluids, 187
 - of heat, 198–202

Stokes's theorem, 229-230
 Strains in elastic solid, 172-183
 Streamlines, 186
 Stresses in elastic solid, 172-183
 String, vibrations, 120-158
 Structure of atoms, 425*ff.*
 Subshells of electrons in atoms, 429
 Sum of states, 468
 Superposition of transient and forced motion, 33-35
 Surface charge density, 212
 Surface forces, 173
 Symmetrical top, 102-104
 Symmetry of stress tensor, 175

T

Temperature dependence, of chemical reactions, 491
 of electron energy in metals, 538-540
 of polarizability and dielectric constant, 555
 Temperature gradient, 197
 Temperature vibrations of a crystal, 474-480
 Thermal conductivity, 197
 effect of electron distribution on, 539
 Thermal equilibrium and canonical ensemble, 457
 Thermal expansion, definition of coefficient, 471
 Thermal pressure in solids, 479
 Thermionic emission, 352-353, 539-540
 Thermodynamics, laws of, 460
 Thomson, G. P., electron diffraction, 336
 Thomson, J. J., scattering of light, 295, 548
 Top, precession, 92, 102-104
 Torque, 92-93
 due to shearing stresses, 174-175
 Torque-free motion of symmetrical rigid body, 98-99
 Total reflection, 265-267
 Transformation, contact and point, 87

Transient flow of heat, 203-208
 Transients, initial conditions for, 29
 Transition probabilities, 393-397
 and selection principles, for atoms, 416
 for molecules, 483
 Transverse waves in elastic solid, 176-178
 Traveling waves, 132, 149
 Tubes of flow, 186
 Turbulent flow, 189
 Two-center problem, wave functions, 527-528
 Two-dimensional oscillator, 84-86
 relation to coupled systems, 114-117
 Types of substances, classification, 447*ff.*

U

Uncertainty principle, 333-339
 relation to phase space, 370-371
 Unit vectors, 48
 in function space, 138
 Units, electrical, 227

V

Valence, homopolar, 442-449
 in hydrogen molecule, 526
 Van der Waals' equation, 464-466
 Van der Waals' force, 440-441, 551-553
 Variable mass, relativistic, 78
 Variable tension and density of string, 146*ff.*
 Variables, separation of, 163
 in quantum theory, 362
 Variation, of constants, method of, 391-392
 of an integral, 339
 Vector, 48-56
 Vector model for angular momentum, 415
 Vector operations in generalized coordinates, 199-201, 229
 Vector potential, 231-232, 241-243
 for oscillating dipole, 289-290

Velocity of light, 243-244, 270ff.
 Velocity field of flowing fluid, 185-186
 Velocity potential, 188ff.
 Vibrations, of coupled systems, 107-118
 of crystals, 474-478
 of elastic solids, 172-180
 of membranes, 160-170
 of molecules, 480-481
 of particles, 19-36
 of strings, 120-158
 Virial coefficients, 464
 Viscosity, 192-194
 Volt-electron, definition, 408

W

Wave equation, for electromagnetic field, 243-244, 253
 for membrane, 164
 in polar coordinates, 171
 in wave mechanics (*see* Schrödinger's equation)
 Wave functions, central-field problem, 418-423
 determinant form, 504-505
 hydrogen, 418-423
 linear oscillator, 357
 two-center problem, 527-528

Wave mechanics, general principles, 335-343
 many-body problem in, 423-433
 one-dimensional motion in, 346-356
 Wave normal, 253-254
 Wave packet, 337-338
 Waves, elastic, 176-180
 electromagnetic (*see* Electromagnetic waves)
 progressive and standing, 149
 on strings, 132, 149-151, 156
 Weighted string, 131, 136
 Wentzel-Kramers-Brillouin method, 347-356
 Work, 41
 as a scalar product, 52
 Work function for metals, 543

X

X-ray series, nomenclature, 425, 437

Y

Young's modulus, 182

Z

Zeeman effect, 556

